NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

DISSERTATION

EXTRACTING VALUE FROM ENSEMBLES FOR CLOUD-FREE FORECASTING

by

Cedrick L. Stubblefield

September 2011

Dissertation Supervisor: Joshua Hacker

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 2011	**3. REPORT TYPE AND DATES COVERED** Dissertation	
4. TITLE AND SUBTITLE: Extracting Value from Ensembles for Cloud-Free Forecasting		**5. FUNDING NUMBERS**	
6. AUTHOR(S) Cedrick L. Stubblefield			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000		**8. PERFORMING ORGANIZATION REPORT NUMBER**	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A		**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol Number ___ N/A____			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution is unlimited		**12b. DISTRIBUTION CODE**	

13. ABSTRACT (maximum 200 words)

The Air Force Weather Agency (AFWA) is currently producing cloud-free forecasts for several agencies, but operational forecasts do not incorporate forecast uncertainty. Uncertainty can be forecasted via an ensemble created with perturbed initial conditions. We combine AFWA's global cloud analysis and cloud advection model with the National Centers for Environmental Prediction's global weather ensemble to study the potential for ensemble cloud-free forecasting in support of space-based image collection. A year of ensemble forecasts forms the evaluation dataset. The operationally relevant cloud-free forecast threshold (cloud cover less than 30%) is evaluated over sets of 24-km grid boxes in three climatologically different regions. The analyses and forecasts favor cloud-cover values near 0% and 100% cloud cover, making skill metrics that assume normal statistics mostly inappropriate. Thus we focus on contingency table metrics at the 30% threshold and argue that the odds ratio is most appropriate. Because costs of satellite image collection are largely unknown or classified, and typical cost/loss models may not apply, we also invoke utility theory to quantify operator benefits obtainable from the ensemble. Ensemble skill is apparent, and utility for risk-averse users in persistently clear, cloudy, and variable regions/seasons yields up to a 20% increase in operational efficiency.

14. SUBJECT TERMS ADVCLD GACE Cloud-Free Forecast Value Utility WWMCA GEFS			**15. NUMBER OF PAGES** 219
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

EXTRACTING VALUE FROM ENSEMBLES FOR CLOUD-FREE FORECASTING

Cedrick L. Stubblefield
Major, United States Air Force
B.S., Florida State University, 1999
M.S., Naval Postgraduate School, 2005

Submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN METEOROLOGY

from the

NAVAL POSTGRADUATE SCHOOL
September 2011

Author: _____
Cedrick L. Stubblefield

Approved by:

Joshua Hacker Professor of Meteorology Dissertation Supervisor	Patrick Harr Professor of Meteorology Dissertation Committee Chair
Karl Pfeiffer Professor of Information Sciences	Philip Durkee Professor of Meteorology
Wendell Nuss Professor of Meteorology	

Approved by: _____
Wendell Nuss, Chair, Department of Meteorology

Approved by: _____
Douglas Moses, Vice Provost for Academic Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Air Force Weather Agency (AFWA) is currently producing cloud-free forecasts for several agencies, but operational forecasts do not incorporate forecast uncertainty. Uncertainty can be forecasted via an ensemble created with perturbed initial conditions. We combine AFWA's global cloud analysis and cloud advection model with the National Centers for Environmental Prediction's global weather ensemble to study the potential for ensemble cloud-free forecasting in support of space-based image collection. A year of ensemble forecasts forms the evaluation dataset. The operationally relevant cloud-free forecast threshold (cloud cover less than 30%) is evaluated over sets of 24-km grid boxes in three climatologically different regions. The analyses and forecasts favor cloud-cover values near 0% and 100% cloud cover, making skill metrics that assume normal statistics mostly inappropriate. Thus we focus on contingency table metrics at the 30% threshold and argue that the odds ratio is most appropriate. Because costs of satellite image collection are largely unknown or classified, and typical cost/loss models may not apply, we also invoke utility theory to quantify operator benefits obtainable from the ensemble. Ensemble skill is apparent, and utility for risk-averse users in persistently clear, cloudy, and variable regions/seasons yields up to a 20% increase in operational efficiency.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

ADVCLD	Advect Cloud
AFW	Air Force Weather
AFWA	Air Force Weather Agency
AVHRR	Advanced Very High resolution Radiometer
CDFSII	Cloud Depiction and Forecast System Version II
CF	Cloud-free
CLM	Cloud mask
CPS	Condensation pressure spread
CTH	Cloud top height
DCF	Diagnostic Cloud Forecast
DMSP	Defense Meteorological Satellite Program
DT	Probability decision threshold
DVote	Democratic vote
ECMWF	European Center for Medium range Weather Forecasting
EMean	Ensemble mean
ETR	Ensemble transform with rescaling
EUMETSAT	European Organization for the Exploitation of Meteorological Satellites
EU	Expected utility
EUV	Expected utility value
EV	Expected value
Fcst	Forecast
FY	Fiscal Year
GACE	Global Advect Cloud Ensemble
GEFS	Global Ensemble Forecast System
GFS	Global Forecast System
HQ AFWA	Headquarters Air Force Weather Agency
HSS	Heidke Skill Score
IR	Infrared
ITCZ	Inter-tropical convergence zone
LEO	Low Earth Orbit
NCEP	National Centers for Environmental Prediction
NH	Northern Hemisphere
OLR	Outgoing longwave radiation
NWP	Numerical Weather Prediction
OLS	Visible and infrared sensors of DMSP
ORSS	Odds Ratio Skill Score
ROC	Receiver Operating Characteristic
SH	Southern Hemisphere
SFCTMP	Surface-temperature model
STD	Standard deviation
STTPS	Stochastic total tendency perturbation scheme

TSS	True Skill Statistic (Score)
URank	Uniform ranks
UTC	Coordinated universal time
VRH	Verification Rank Histogram
WRank	Weighted ranks
WWMCA	World Wide Merged Cloud Analysis

ACKNOWLEDGMENTS

To the only wise God be glory forever through Jesus Christ (Romans 16:27 NIV)

My thanks go to wife and best friend, Tamico, who is more than I deserve. She supports me in every endeavor and often puts her desires on hold to ensure that I can focus on my life's pursuits. My sons, Cedrick and Kyle, and my baby boy, Sean, are an ever-present inspiration. My parents, Pastor Arnold and Ruby Stubblefield, instilled in me strong values and the confidence to believe that integrity and hard work are enough to bring dreams to reality. My wife's parents, Pastor James and Jewanta Jamison, who support me as though I am their only child—"in-laws" is an inappropriate name for them.

Special thanks to Dr. Joshua Hacker and the rest of my dissertation committee. Dr. Hacker is a brilliant professor who helped me maintain a research approach that was clear and academically rigorous. Dr. Karl Pfeiffer's military perspective, experience, and sharp focus were a daily source of motivation. Thanks to Dr. Patrick Harr, Dr. Wendell Nuss, and Dr. Philip Durkee for maintaining an open door policy and assisting me in overcoming my most difficult research challenges.

Thanks go to Mr. Robert Creasey for expediting my research by teaching me scripting and data management techniques. Thanks to my student colleagues and the staff of the meteorology department who helped me in countless ways.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Forecasts of cloud-free conditions are of particular interest to seismologists, oceanographers, geologists, intelligence operators, and military strategists who depend on space-based satellite imagery to obtain global, regional, and local information. While most can leverage geostationary satellites to obtain cloud-free imagery, foreshortening (distortion of image not parallel to viewing plane) limits the accuracy of radiance retrieval above 60N and 60S. Geostationary satellite resolution ranges from 1 km to 8 km, which is rarely sufficient for pinpoint intelligence and/or military operators. Low Earth Orbit (LEO) satellites are better suited to support intelligence collections because pixel sizes are on the order of 0.5 km, but revisit times are limited by the number of LEO satellites in orbit. When the goal is obtaining high resolution, cloud-free imagery of a target, but continuous surveillance of a point or location of interest is not available, target selection becomes critical. Furthermore, cloud-free forecasts are essential input in developing image collection prioritizations for LEOs with an adjustable field of view.

Air Force Weather (AFW) is currently producing automated and manual cloud-free forecasts for the U.S. Department of Defense, intelligence agencies, military surveillance, and reconnaissance support, but their forecasts do not incorporate the uncertainty, or confidence, in the accuracy of cloud cover predictions. Forecasts are generated based on subjective analyses of available satellite imagery and modeled cloud data over regions of interest. The forecasts provide the user with information about the percentage of the region that will be cloud-free at the time of image retrieval with a 3–24 h lead-time. The research herein is a first step towards using an ensemble-based method in the production of cloud-free forecasts.

This effort represents the first quantification of global probabilistic cloud-cover forecast skill and novel application of utility theory in assessing the value of cloud-cover predictions in cloud-free forecast operations. We combine AFWA's cloud advection algorithm and analysis with a global ensemble weather prediction system to produce a global cloud-forecast ensemble. In doing so, we intend to show the potential for ensemble predictions to demonstrate skill and provide value in support of cloud-free

collection operations. Unlike the current method, ensemble forecasts provide a dynamic and objective way of articulating uncertainty of a forecast.

We interpret the ensemble forecast in four primary ways, deterministic (control and mean forecast) and probabilistic (democratic voting, uniform ranks and weighted ranks). Each forecast interpretation or ensemble forecast method is evaluated regionally for skill. The skill of the ensemble is calculated via common deterministic and probabilistic measures and evaluated by region, cloud cover frequency, and probability decision threshold.

The usefulness of the forecast is defined by the extent in which the forecast adds value to operational outcomes. Without forecast value information, a decision maker cannot optimize decision processes. Research studies have shown that decisions made based on cost/loss ratio information can improve the outcomes of risk management exercises. However, cost and loss information is often uncalculated, miscalculated, or subject to security clearance restrictions. Although the cost/loss model provides a simple means of calculating the value of forecasts, the complexity in defining the cost/loss ratio makes it less than practical for operational environments.

We introduce utility theory to cloud forecasting and assess the expected value of ensemble forecast information in cloud-free image collection operations. Utility theory has been around since the 1700s, but has been primarily cast in an economic context. The cost/loss ratio has been most often used to define the utility of forecasts, but utility can also be defined without the explicit inclusion of economic information. We use this scarcely employed approach to calculate the value of weather forecasts to evaluate whether the skill of the ensemble translates to tangible or intangible (though quantifiable) gains for the decision maker.

II. BACKGROUND

A. PREDICTABILITY

Lorenz (1963) demonstrated the idea that our ability to accurately predict future states of the atmosphere is limited by inaccurate observations and inaccurate boundary conditions. The inherent error in the initial condition precludes a perfect representation of a future state of the atmosphere. Furthermore, the ability to predict future states of the atmosphere can be defined by three categories of error growth (Lorenz 1969) as follows:

- *At all future times the error remains comparable to or smaller than the initial error. The error may be kept arbitrarily small by making initial error sufficiently small.*
- *The error eventually becomes much larger than the initial error. At any particular future time the error may be made arbitrarily small by making the initial error sufficiently small, but no matter how small the initial error (if not zero), the error becomes large in the sufficiently distant future.*
- *The error eventually becomes much larger than the initial error. For any particular future time there is a limit below which the error cannot be reduced, no matter how small the initial error (if not zero) is made.*

Category one forecasts have an infinite range of predictability; category two forecasts have limited predictability but can be improved by decreasing initial error; and category three forecasts have an inherent limit to predictability that cannot be improved beyond some point by reducing initial error. The atmosphere falls into category three because of the non-linearities appearing in the equations of motion.

Cloud forecasts have a finite range of predictability due to error growth resulting from non-linear interactions between clouds and the environment and non-linear errors in predicting the advecting winds. Cloud evolution (development, movement, and dissipation) is too sensitive to environmental influences to maintain accuracy over long periods of time. We can expect that synoptic scale cloud features are more predictable than mesoscale cloud features, which are predictable on the order of 2–6 h (Droegemeier 1990). Different time scales of horizontal advection, vertical advection, and entrainment in cloud development contribute to temporal variations in error growth. Predictability not

only changes with horizontal and temporal scales but is also dependent on latitude (lower in tropics), season (higher in NH winter), and synoptic system (Chou 1989).

Predictability expresses the need for what Tennekes et al. (1986) terms a "forecast of forecast skill." The predictability of an event tends to vary based on its temporal proximity to observations, the number of times the event has been observed, and the understanding of the triggers and variables that impact the event. Therefore, some events can be more predictable than others—rare events can be less predictable than events that are well documented. Consider sunrise, consistent observations, frequent experiences, and scientific knowledge of the Earth's orbit makes it easy to predict dawn or dusk. If we are more interested, however, in forecasting whether cloud cover will inhibit locals from seeing the sunrise than we are about predicting the sunrise itself, cloud-cover conditions tomorrow can be more predictable than conditions six months from now. The further displaced an atmospheric forecast of an event is from current observations the less predictable the event becomes. This is due in large part to atmospheric instabilities, non-linear processes, and the structure of imperfections in the initial conditions (Palmer and Hagedorm 2006).

Information about the amount of cloud covering specific locations on the earth's surface at a given time is critical to a number of civil and military operations. The ability, however, to accurately quantify atmospheric cloud cover is limited by uncertainty, manifested as errors in the cloud development predictors: moisture, synoptic-scale winds, and rainfall (Roach 1994). If these parameters are not characterized correctly, inaccurate cloud development and displacement will ensue.

B. PROBABILITY THEORY

Frequency theory is the basis for probability calculations within this research though several probability theories have been examined in an effort to quantify uncertainty. We find it beneficial to introduce the three major theories: frequency, logical, and subjective. These theories follow from the same axioms (Wilks 2006):

1) $$0 < p(E_i) > 1 \qquad i = 1,2,3...n$$

2) $$\sum_{i=1}^{n} p(E_i) = 1$$, for compound events

3) $$p\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} p(E_i)$$, for mutually exclusive events

Axiom 1) states that the probability of event E occurring is nonnegative and less than or equal to 1. Events with probability 0 (or 1) are expected not to occur (or to occur) with 100% certainty. Axiom 2) states that probability of each event in the sample space sums to 1. Axiom 3) states that the probability of an event occurring from the union of disjoint events is the sum of their individual probabilities.

"The frequency theory defines the probability of an outcome as the limiting frequency with which that outcome appears in a long series of similar events" (Gillies 2000). In short, it is the ratio of time an event happens to the number of times it could have happened. Because of its usefulness in estimating probabilities, frequency theory has become the mainstream method of defining probability (Wilks 2006). Users of frequency probability, however, must assume that past events are statistically reliable and unbiased so that they represent future events.

Some have questioned the meaningfulness of frequency theory, which is limited by sampling error (Lee 1971). If an event occurs 200 times out of a possible 1000, then the probability of occurrence is 20%. If the event is a fair coin toss producing heads, the estimated probability does not reflect the true proportion of 50%. This highlights the finite problem of frequency theory. The next 1000 coin flips could produce 800 heads making the frequency over the long run equal the true proportion of 50%. This is also the case if heads comes up 500 out of the next 1000 times. Frequency probability is limited by the finite number of cases used to define it. Therefore, it tells us when an event is most likely to happen and much less the confidence in the occurrence of the event (Macmillan 2002).

Another problem with frequency probability is that it uses a set of events to define a single event (Audi 1999). Lee (1971) uses a 3/5 probability of rain forecast in San

Francisco to illustrate this point. A 3/5 probability of rain refers to a particular time (tomorrow) and place (San Francisco). Thus, the occurrence of three rainy days out of five cannot be verified on a single day. This is partly the reason meteorologist turn to Bayesian interpretation of probability.

Bayesian interpretation maintains a "history is prolog" view of frequency probability. History states that rainy days in San Francisco occurred 3/5 times when atmospheric conditions were like they are today, but this too is problematic. Non-linear characteristics of the atmosphere and unaccounted for variability between the five previous events and today can impact the probability of rain tomorrow.

Logical and subjective probability theories are synonymous with objective and "personal belief" interpretations of Bayesian probability theory. Logical (frequency) probability theory assumes that the statistical occurrences of past events are reliable and unbiased guides for future events (Davidson 1991). The logical interpretation of probability proposes that event E is always related to some evidence F and has a probability p(E|F) (Lindley 1985). This interpretation of probability is often preferred for its sensitivity and adaptability to new information.

The personal probability viewpoint of Bayesian probability theory suggests that a person's level of belief defines the probability of event occurrence. The term "subjective probability" is most often preferred over "personal probability" because the probability is not really personal but based on an ideal person (Lee 1971). Subjective probability is similar to logical probability in that probability definitions are related to evidence F. The difference is logical theory says that given p(F) only one p(E) exists, but different people can have differing p(E) given the same p(F). It would be an interesting exercise to examine the improvement in cloud-free forecast skill or value if probability forecasts were conditioned by the probability of a sister event (i.e,. frontal passage, shift, long wave pattern), but the broad scope of our research does not lend itself to the Bayesian interpretation of probability.

We elect, instead, to base probability forecasts on frequency theory and perform the first comprehensive analysis of probabilistic cloud-free ensemble forecasts. We focus

on the current skill and value of predicting cloud cover in general rather than the results of a probability forecast given the probability of a precursor event.

C. ENSEMBLE FORECASTING

Leith (1974), the first to introduce the concept of ensemble forecasting, showed that a 10-member ensemble forecast can improve predictions with lead-times of 10 days (Sivillo et al. 1997). Leith demonstrated that running a numerical weather prediction (NWP) model several times with differing or perturbed initial conditions, which well represented the uncertainty in the analysis, could improve 6–10 day forecasts. It would take an infinite number of model runs with a perfect model to produce a perfect ensemble forecast. Neither is possible.

Leith's approach to ensemble forecasting, referred to as "Monte Carlo forecasting," chooses perturbations that have horizontal and vertical structures similar to the forecast errors compatible with the typical uncertainty of the analysis (Leith 1974). This parameterized method does not reflect the true structure of uncertainty in initial conditions (Toth and Kalnay 1993). In an attempt to reflect the real initial uncertainty, Mullen and Baumhefner (1987) adjusted the Monte Carlo forecasting method to include an estimation of analysis errors.

By 1992, the maturity of the science and increase computer processing capability made operational ensemble forecasting possible at the National Centers for Environmental Prediction (NCEP). Toth and Kalnay (1993) demonstrated that random perturbations do not account for fast-growing errors in the analysis, which are important in capturing the true evolution of the atmosphere. To account for these fast-growing errors a "breeding" method was devised. The method models how fast-growing errors are "bred" into the analysis cycle (Toth and Kalnay 1993). Other methods that have been tried include time-lagged (Hoffman and Kalnay 1983) and singular vector (Ehrendorfer et. al. 1998) approaches.

According to Kalnay (2003), ensemble forecasting is important in two primary ways. 1) The predictions of the mean forecast averages out the most uncertain components of the forecast, which produces better results than deterministic forecasts

after a few day. 2) Using the ensemble probabilistically provides a means to evaluate the reliability of the forecast. She showed that the ensemble mean forecast demonstrated more skill than the control forecast beyond five days. No significant differences are apparent, however, at shorter lead-times. Hence, minimal differences are expected between the ensemble mean and control forecasts at the short lead-times required for cloud-free image collections.

Ensemble forecasts have been applied in four primary ways (Anderson 1996): 1) The deterministic forecast is replaced with the ensemble average (mean) forecast, 2) The deterministic forecast is replaced with the mean of each distinct ensemble forecast cluster which represents the different forecast states of the ensemble, 3) The ensemble spread (difference between ensemble-member forecasts) is used to infer skill, and 4) Forecasts are calculated from the probability distribution of the ensemble. We will examine applications 1 and 4, but will omit 2 and 3. The bi-modal distribution of cloud cover analyses and forecasts—discussed in greater detail later—may make clustering and spread statistics less meaningful than traditional metrics.

Limitations in computational resources have forced previous probabilistic cloud forecast studies to avoid the development and applications of cloud ensembles. Weymouth et al. (2007) used Bayesian Network techniques to estimate the uncertainty in forecasting fog and low clouds for major airports in Australia. They developed a network structure that improved forecast skill significantly over previously used guidance at some locations. Kemp and Allis (2007) combined numerical weather prediction data from the Weather Research and Forecasting model, objective cloud analyses from the Cloud Mask Generator developed by TASC, and a logistic regression to derive cloud fraction forecasts that can be interpreted as cloud probability forecasts. In both cases, probability forecasts were produced using moisture as a surrogate for and predictor (along with u and v winds) of clouds. These studies showed that understanding and communicating uncertainty in cloud forecasts can improve forecast skill and value.

Unlike statistical methods, ensemble forecasts provide a dynamical and flow-dependent method of quantifying and communicating forecast uncertainty. Ensemble forecasts provide a simple method to mitigate the non-linear nature of environmental

prediction by producing multiple forecasts with varying initial condition. Perturbations in the initial conditions serve as an estimate of the analysis probability density function and should be consistent with analysis errors. Properly defining the initial conditions can limit the impact of inherent errors in predicting future states of the atmosphere.

D. ECONOMIC VALUE

Economic value—we use this term interchangeably with monetary value—is one of several methods used to measure the effect of forecasts on decision process. Economic value is particularly applicable when a decision maker must make choices between dichotomous events that have significant financial impacts. Forecasters should avoid the temptation to make skill and accuracy the focal points of forecast verification; the value or usefulness of the forecast to its users is the key measure of forecast performance. Forecasts can show significant skill and accuracy, but provide little value to a given decision process (Murphy and Ehrendorfer 1987; Murphy 1997). Unique sensitivities to operational costs and potential losses due to following the forecast can make the value of the forecast highly variable from user to user.

A simple method used to capture the user dependent value of forecasts, in economic terms, is the cost/loss ratio (Wilks 2001). The cost-loss ratio is a decision analysis tool that helps quantify the economic consequence of a decision by comparing the difference between the cost of protecting assets against an adverse event and loss incurred by not protecting (Thompson 1952) assets. Users are expected to objectively act based on his/her organization's sensitivity to a given phenomenon.

Table 1. 2x2 Economic Contingency Table: Forecast and Observation

		FORECAST/ACT		
		YES	NO	Total
OBSERVATION	YES	Hit (HT) (Cost)	Miss (MS) (Loss)	Total Yes Obs (TYobs)
	NO	False Alarm (FA) (Cost)	Correct Rejection (CR) (No Cost)	Total No Obs (TNobs)

Table 1 shows the four typical economical expense categories relative to forecast outcomes. The table provides a simple method of capturing the operational value of dichotomous decisions (e.g., yes/no, go/no-go). A cost C is incurred whenever the user decides to take protective action against an adverse weather event. This cost is paid regardless of whether the event actually occurs or not. A loss L is incurred when the adverse weather event occurs and no protective action is taken.

Consider a hypothetical situation where a user never deviates from hail forecasts. A cost C is incurred whenever hail is forecasted (aircraft are sheltered) and a loss L is incurred when hail is not forecasted but observed (aircraft damaged). When hail is neither forecasted nor observed, no cost or loss is incurred.

Cost/lost information, in conjunction with the probability of event occurrence, can be used to optimize a decision process. The operator who decides to act in response to information that the event occurs with frequency or probability *p* will incur a cost C each time aircraft are sheltered. When the operator decides not to protect, the operator will suffer loss L in damaged aircraft at probability *p*. This means the operator should shelter aircraft whenever the cost is less than the probability of loss or the cost/loss ratio is less than the probability of event occurrence:

$$C < pL \qquad 1$$

Or

$$\frac{C}{L} < p \qquad 2$$

The equation highlights the utility of forecast value over skill and accuracy. If cost C of protecting is larger than the expected loss L in not protecting, operations will be optimized by ignoring the forecast—even the perfect forecast. When C<0, the goodness of the forecast also becomes a moot point because the operator will always protect. Thus, economic value considerations are relevant when 0<C/L<1. Following Wilks (2001), we can define:

$$EE_f = \left(\frac{HT + FA}{n}\right)C + \left(\frac{MS}{n}\right)L \qquad 3$$

$$EE_c = \min(nC, n\bar{o}L) \quad (4)$$

$$EE_p = \bar{o}C \quad (5)$$

$$\bar{o} = \left(\frac{HT + MS}{n}\right) \quad (6)$$

The expected expense of the forecast (EE_f) is the total cost of protecting and loss of not protecting resulting from exclusively following the forecast. The expected expense of following climo (EE_c) is the minimum between the cost of always following the forecast and never following the forecast. The expected expense of the perfect forecast (EE_p) represents the cost of acting only when event is observed. The climatological probability of the event occurring is $\bar{o}$ and n is the total possible cases

Several factors, however, preclude the use of cost/loss information in measuring the operational value of the forecast. There is no direct cost in capturing images via most space-based platforms. The advent of digital imagery has eradicated this expense. It remains that the quantifiable costs of image collection results from initial satellite launch and subsequent image processing. The financial impacts of these pre- and post-collection operations are either unavailable or classified. Furthermore, attempting to quantify the loss due to missing a collection opportunity is not practical. Thus, it is convenient to employ utility theory, which allows us to subjectively quantify the operational value of a forecast in the absence of valid cost/loss information.

E. UTILITY THEORY

1. Definition

Although economic value calculations are beneficial as budgeting tools, ballpark figures, briefings, and long term planning, the value of military operations is measured by mission success and failure. The economic impact of losing an aircraft because of adverse weather conditions cannot communicate unit degradation in strategic and tactical readiness. How does a commander put a price tag on the potential lives lost when protective actions are not taken during an adverse weather event? The value or worth of a bombing mission is not wholly measured in dollars and cents. There are a myriad of combat and no-combat operations where economic value falls short in providing

sufficient information required for proper risk assessments. The value of these types of missions is subjectively centered on the level of task completion and measured by the level of user satisfaction. The major challenge in capturing the true value of an operational forecast is that forecast value is a complex function of several variables that are difficult, if not impossible, to quantify and highly user/operation dependent.

Unlike economic value, which only varies with fluctuations in operational costs, utility value accounts for fluctuations in the operator's sensitivity to mission outcomes. The cost or loss associated with performing a given mission may not change for years, however, the perceived value of mission success and failure can change with every supervening operation. The cost of collecting and processing an image may be the same today as it was two years ago. But, if the image was not successfully collected over the two year period, the desire to collect the image can increase significantly because of unfulfilled intelligence requirements. This abstract way of defining worth is subjective and difficult to measure; but in an effort to quantify it we turn to Utility Theory.

Lee (1971) concludes, "predictions based on the utilities are typically superior, but only slightly so." The fact that utility value is only "slightly" better makes it less attractive than simpler techniques. This could be the reason utility is generally neglected, but we welcome the idea that utility value can be used interchangeably with economic value, for we seek to replace the expected value (EV) that arises from expense with expected value that arises from expected utility (EU).

Bentham (1823) defined utility as "that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness." He further states that utility describes a person's desire to avoid pain and the measure of pain is regulated by the intensity, duration, certainty, proximity, and extent of the circumstance. A person finds the most gratification in correctly protecting against an adverse event when the level of devastation (intensity) or length of recovery (duration) are high. The appreciation of a forecast is also high when events are rare (certainty), not easily accessible (proximity), or affect a large number of people (extent). The latter three are clearly ascribable to collection operations.

The utility of a forecast is directly related to the certainty or uncertainty of clear conditions relative to the atmospheric dynamics of a given region. Proximity or convenience of collection operations is low because polar orbiting satellites do not provide access to targets 100% of the time. The consumers of intelligence information are numerous. The success or failure of a given collection mission impacts civilian, government, and military strategic and tactical operations and personnel.

Utility is individually specific and can be temporally bounded. It cannot be deemed correct or incorrect because it is defined by the user's perception of the consequences. This approach to assessing value generates potential limitations, which were outlined by Lee (1971). First, the assigned utility may not match the realized utility of the outcome. No one person has the amount of experience necessary to adequately evaluate outcomes. Second, utility can be counter intuitive or over dependent on the user. Two commanders facing the same scenario may perceive them differently. One commander may find an operation designed to gain a tactical advantage of great utility value to the overall mission of the unit. Another may consider the risk to life and/or future operations too high and consider the utility value of the operation to be low. Third, psychological factors are not considered. Be it evolutional or environmental, people change. An operator who assigns a utility value to an outcome today may not assign the same utility value in the future, even though conditions are the same.

EV and EU are defined as:

$$EV = \sum_{i=1}^{n} p_i v_i \qquad 7$$

and

$$EU = \sum_{i=1}^{n} p_i u_i \qquad 8$$

where EV_i and EU_i, in probability theory, is the weighted average or payoff of n mutually exclusive cases i where payoffs v and u occur at respective probability p. Note that EV and EU are calculated in the same manner except concrete values such as dollars in EV are replaced with subjective utility values in EU. To appreciate the difference between

EU and *EV*, we will use the gambling problem that is often used to demonstrate the difference. Here we follow the example put forth by Lindley (1985) and consider a gambler who is willing to gamble \$300 for a chance to win \$1,000. The gambler is offered the following three prizes:

v=\$1000	p = .20
v=\$500	p = .40
v=\$0	p = .40

The expected value of the gamble is

$$EV = .2(1000) + .4(500) = \$400$$

EV suggests that the gambler should place the bet. However, utility theory necessitates the consideration of the relative value, or gambler satisfaction, in winning the \$1000 and \$500 prizes. Since the range of our outcomes is \$1,000 and \$0, we will arbitrarily assign utility values of 1 to the maximum amount (\$1000) and 0 to the minimum amount (\$0). We then ask 2 gamblers to assign utility values for winning \$500 and keeping the \$300. If gambler 1 is hoping to earn enough money to go on a summer trip that costs \$900, this person would see \$1000 as having more than twice the utility of \$500 and might assign a utility value of .2 to the \$500 prize. Gambler 1, who has limited resources, might not be inclined to risk losing everything and assign a utility value of .6 to keeping the \$300. Gambler 2, more affluent than gambler 1, may be indifferent to winning either of the prizes and losing the \$300. Because winning the larger amount is preferred regardless of economic status, gambler 2 assigns a utility value of .9 to \$500 and .1 to keeping the \$300. Our expected utility for each gambler is as follows:

Gambler 1: $EU = .2u(\$1000) + .4u(\$500) = .2(1) + .4(.2) = .28$

Gambler 2: $EU = .2u(\$1000) + .4u(\$500) = .2(1) + .4(.9) = .38$

Gambler 1 has an expected utility of .28 which is less than his selected utility of keeping the \$300 (.6). Gambler 2 has an expected utility of .38 which is more than her selected utility of keeping the \$300 (.1). Utility theory suggests that Gambler 1 keep the \$300 and Gambler 2 take the bet. The fact that each gambler, considering the same bet, is advised to take a different action highlights the key difference between expected value

and expected utility. All gamblers share the same expected (economic) value, but expected utility—to what extent the gamble is advantageous—depends on perspective.

Utility Theory, says Beach (2005), "is used to represent preferences among potential outcomes of a decision." Three common assumptions are used when assigning utility values. These assumptions are listed below and demonstrated with alphanumeric outcomes:

Assumption 1: Preference or indifference exist between outcomes

$A \succ B$ or $A \sim B$

Assumption 2: Outcomes preferences are transitive

If $A \succ B$ and $B \succ C$, then $A \succ C$

Assumption 3: A preference is greater than any component part

If $A = A1 + A2$, then $A \succ A1$ and $A \succ A2$

For example, 1) a decision maker may prefer to watch a movie rather than a theatrical play. The play costs more than the movie, but the movie is valued more by the decision maker. 2) The decision maker may rather see the play than attend a sporting event. It should follow that the decision maker also prefers the movie over the sporting event. 3) If the movie ticket comes with popcorn, neither the popcorn nor the movie ticket can be valued more than the movie ticket and popcorn together, assuming there are no other economic factors to consider such as sales and discounts.

2. Applications

Table 2 contains a list of hypothetical probabilities and utility values that could be associated with the outcomes of a space based collection operation. There are two collection decisions to choose from, collect (d_1) and not collect (d_2). There are two possible cloud-cover conditions of the desired target, cloudy (q_1) and clear (q_2); clear is good and cloud is bad. The variable p_1 is the probability of clear and p_2 (1 - p_1) is the probability of cloudy. The utilities assigned correspond to the outcomes in Table 1. The details of the selection method and application of these utility values are saved for

Chapter III. We intend here to simply orient ourselves with calculating expected utility of dichotomous decisions and using expected utility to evaluate the expected value of ensemble forecasts in cloud-free collection operations.

Table 2. Decision table for the cloud-free collection problem with numerical values for the utilities and probabilities (After Lindley).

	d_1: Collect	d_2: Don't Collect	$p_j=p(q_j)$
q_1: Good (clear image)	1	0	.3
q_2: Bad (cloudy image)	0	.8	.7

Using the decision table (Table 2), we can now calculate the expected utility for a user with the given utility values,

$$EU(d_1) = u_{11}p_1 + u_{12}p_2 = (1)(.3) + (0)(.7) = .3$$

$$EU(d_2) = u_{21}p_1 + u_{22}p_2 = (0)(.3) + (.8)(.7) = .56$$

The operator is expected to choose the decision which has the largest expected utility. In our example, decision makers should choose d_2 over d_1 because of the high probability of collecting a cloudy image makes not collecting the maximum expected utility. If the probabilities were flipped (p_1=.7; p_2=.3), the expected utility for d_1 and d_2 would be .7 and .24 respectively. Therefore the user should elect to collect (d_1) in the area with a high probability of collecting clear imagery. Maximum expected utility is denoted as,

$$\max_i \sum_{j=1}^{n} u_{ij} p_j \qquad 9$$

where $\max_i$ is the maximum expected utility which corresponds to decision d_i.

If the desired collection is in a persistently cloudy region, p_1=.3 as before, would the user never choose to collect? Most military operations are sensitive to external variables which introduce additional risks, but never performing an operation simply because of the risk evolved is never an option. What if the decision maker had additional information? A gambler who has a 30% chance of winning would be more likely to

place the bet if supplemental information increased his chances of winning to 70%. Therefore, the remaining challenge is to increase the probability of success by interjecting additional information about the state of the problem. An increase in expected utility, however, is limited to the level of uncertainty that can be removed from the decision.

Expected utility is maximized with the acquisition of perfect information. With perfect information, referring to Table 2, a user would always select decision d_1 when condition q_1 occurs and d_2 when condition q_2 occurs. The expected utility of perfect information results from multiplying the maximum utility values u_j of each decision d_i by the probabilities p_j that correspond to the maximum u_j and summing. The use of perfect information results in an expected utility of .86. Perfect utility can be represented as,

$$\sum_{j=1}^{n} \max_{i}(u_{ij})\, p_j \qquad 10$$

The numerical difference between maximum and perfect expected utility is the expected value of perfect information (Lindley, 1971).

$$EUV_p = \sum_{j=1}^{n} \max_{i}(u_{ij})\, p_j - \max_{i} \sum_{j=1}^{n} (u_{ij})\, p_j \qquad 11$$

Expected value of perfect information represents the maximum gain in utility that can be realized by introducing additional information into the decision process. Continuing with our example, the expected value of the perfect forecast is .03. Forecasts are expected to be perfect, or close to it, when uncertainty is low (probability of event occurrence is 0% or 100%). Very little uncertainty exists in a region that is clear 100% of the time. Therefore, the need for additional information in the form of a cloud forecast is infinitesimal. The same is true for a persistently cloudy region. Additional information provides less value in highly predictable situations. Unfortunately, most decisions are riddled with uncertainty and forecasts are less than perfect. Forecasts are expected to provide the most value when there is a reasonable amount of uncertainty in event occurrence.

Perfect forecasts will naturally provide more utility than uncertain forecasts. However, expected utility for any decision problem is expected to increase with the

consideration of relevant information—even if the information is not perfect (Lindley 1971). If a forecast E is related to an event q_j and the forecast is reliable, we can calculate and use a likelihood function $p(\mathrm{E}|q_j)$ to determine how the operator should act given E.

Table 3. EDecision table for the cloud-free collection problem with imperfect information (After Lindley 1985).

	d_1: Collect	d_2: Don't Collect	$p_j=p(q_j)$	$p(E_1\|q_j)$	$p(E_2\|q_j)$
q_1: Good	1	0	.3	¾	1/4
q_2: Bad	0	.8	.7	¼	3/4
E_1	**.225**	.14			
E_2	.075	**.42**			

Consider a reliable model that provides correct forecasts 75% of the time. This means the probability of the model correctly forecasting clear and cloudy can be represented by $p(E_1|q_1)$ and $p(E_2|q_2)$, respectively. It also means that the probability of the model to incorrectly forecast clear and cloudy conditions (25%) is $p(E_2|q_1)$ and $p(E_1|q_2)$, respectively. With decision d_i and forecast E, we can calculate the model's expected utility.

$$\sum_{j=1}^{n} p(E \mid q_j)\bar{u}_i \qquad 12$$

The expected utility of the model E that results from the Table 2 example is listed in Table 3. The utility for decision d_i, probability p_j, and $p(\mathrm{E}|q_j)$ are multiplied for each decision and summed as outlined by Lindley (1985).

$$d_1E_1 = \left(1 \times \frac{3}{10} \times \frac{3}{4}\right) + \left(0 \times \frac{7}{10} \times \frac{1}{4}\right) = .225$$
$$d_2E_1 = \left(0 \times \frac{3}{10} \times \frac{3}{4}\right) + \left(\frac{8}{10} \times \frac{7}{10} \times \frac{1}{4}\right) = .14$$
$$d_1E_2 = \left(1 \times \frac{3}{10} \times \frac{1}{4}\right) + \left(0 \times \frac{7}{10} \times \frac{3}{4}\right) = .075$$
$$d_2E_2 = \left(0 \times \frac{3}{10} \times \frac{1}{4}\right) + \left(\frac{8}{10} \times \frac{7}{10} \times \frac{3}{4}\right) = .42$$

From Table 3, the largest E value between decision d_1 and d_2, that is within each row, is selected (**bold**) and summed. The expected value of the forecast then becomes the difference between expected utility of imperfect forecast information and maximum expected utility of the operation.

$$EUV_f = \sum_E \max_i \sum_{j=1}^{n} p(E \mid q_j)\overline{u}_i - \max_i \sum_{j=1}^{n} \overline{u}_i \quad . \qquad 13$$

Recall that the maximum expected utility and expected utility with perfect information are .56 and .86 respectively. With model information, the expected utility is .65 which equates to an expected value of .09. It may be helpful to consider this number in terms of money.

Let us say a successful operation is worth $1 million. Multiplying the expected value of forecast information and the economic value of the operation, the user should be willing to pay $90,000 for the model forecast. This is almost 1/10th the value of perfect information but still presents the possibility of improving the decision process.

F. UTILITY FUNCTIONS

Users can have differing sensitivities, or attitudes, toward the risk of collecting cloud filled imagery. In the previous examples we have assumed that decision maker satisfaction follows a linear function. In practice, however, we may find that the function is altogether different. Some utility functions that may apply to this decision problem are non-linear as with non-linearities seen in economic utility, first recognized by Bernoulli (as cited by Shorr 1966).

Figure 1 demonstrates four utility curves that may be applicable to cloud-free forecast operations. The first curve, the risk-neutral curve, suggests that dissatisfaction with cloudy imagery is neutral across the user population. Here the dissatisfaction of a user who assigns a utility value of .8 is twice that of a user that assigns a value of .4. The second curve, the risk lover, implies that the difference between high tolerant or risk prone users is small and most users are willing to accept significant risk to collect a clear image. The third curve, risk-averse, implies that risk prone users can have a relatively high satisfaction with correct rejections, which reduces the difference between high and low tolerant users. The forth curve, the sigmoid utility, is a combination of risk lover and risk-averse curves. Users within the convex section of the curve are risk seeker, and users in the concave section of the curve are risk avoiders.

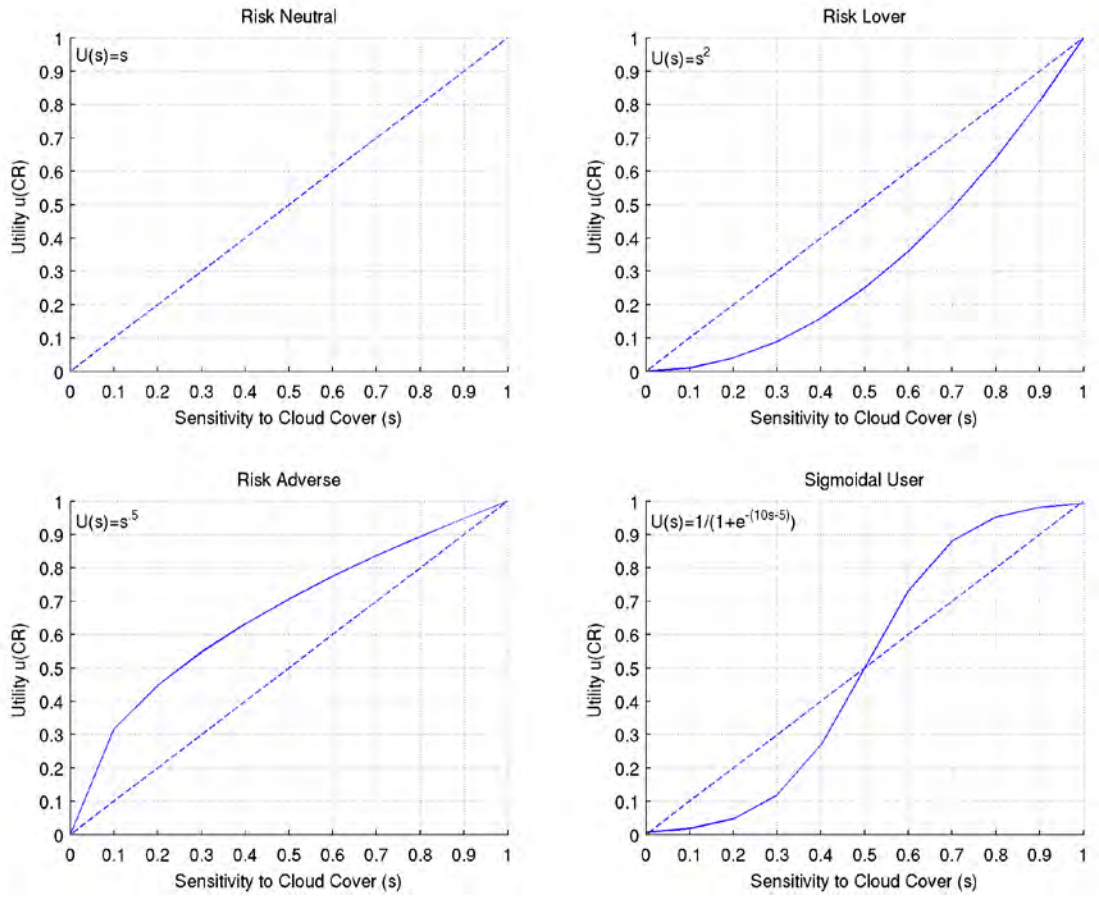


Figure 1. Utility curves represent the risk tolerance or operator sensitivity to cloud cover imagery. Risk-neutral users (top left) are indifferent to cloud cover. Risk lovers (top right) are prone to accept a significant amount of risk with the possibility to collect a clear image. Risk-adverse uses (bottom left) prefer to have information in the mitigation of risk. Users defined by a sigmoidal curve (bottom right) are risk lovers with high priority targets and averse with lower priority targets (After Lawrence 1999)

The risk lover and sigmoid curves are mentioned here only for completion. Rational users are presumed to be either risk-neutral or risk-averse (Lawrence, 1999). Therefore, we limit our utility analyses to forecasts that support users that risk-neutral and risk-averse. We use a linear function to represent the utility of risk-neutral users. This utility function follows to the decision maker's sensitivity to cloud filled imagery.

$$u(CR) = s \tag{14}$$

$$u''(CR) = 0 \tag{15}$$

where *u(CR)* is the utility of correct rejections and *s* is the user's sensitivity to correct rejections. Several utility functions that have been used to describe risk-averse users (e.g., ln(x)), but we choose to utilize the exponent found by Galanter in 1962 (as cited by Lee 1971):

$$u(CR) = s^{0.5} \tag{16}$$

$$u''(CR) = -.25s^{-1.5} < 0 \tag{17}$$

The second derivative of equation 14 is equal to zero so the curve is linear. The second derivative of equation 16 is negative; so the utility function is strictly concave, which is a characteristic of risk-averse curves (Lawrence 1999).

III. DATA

Data from the Air Force Weather Agency and the National Centers for Environmental Prediction are used to develop a global cloud forecast ensemble, which is used to evaluate the skill and value of ensemble forecasts in cloud-free forecast operations. Initialization and verification cloud fields from the World Wide Merged Cloud Analysis (WWMCA) are retrieved from the Air Force Weather Agency in flat file format. The agency maintains an archive of their global cloud analysis data from 2005 to current. Total cloud amount, cloud layer percentage, cloud layer base heights, cloud layer top heights, and cloud mean pixel time are grouped by date and time, which amounts to about 1.6 gigabits per day (uncompressed).

Forecast parameters are retrieved from the National Centers for Environmental Prediction ensemble, the Global Ensemble Forecast System (GEFS). These data are archived locally at the Naval Postgraduate School. Archiving of these data coincide with the beginning of this research, January 2010. Preliminary tests were accomplished with the January data, but the research presented begins in February, the first complete month of archived data. These files are compressed in grib1 format and received daily via file transfer protocol.

A. WORLD WIDE MERGED CLOUD ANALYSIS

The global coverage of WWMCA provides tremendous utility as a global cloud forecast initial condition and observation for verification. The global coverage of the analysis makes it unique among other resources. Other satellite based cloud analyses are limited to hemispheric or regional cloud coverage.

The World Wide Merged Cloud Analysis is produced in the cloud depiction segment of the Cloud Depiction and Forecast System Version II (CDFSII). It results from merging geostationary and polar satellite data; surface observations; and relative humidity, temperature, and wind data from the National Centers for Environmental Prediction's Global Forecast System (GFS). The data is interpolated to a whole-mesh grid (381km at 60 degrees latitude). Higher resolution grids are defined as a fraction of

the whole-mesh grid (half-mesh, 8th mesh, 16th mesh). The name of these nested grids reflects the number of sub-grid cells per a whole-mesh grid (8th mesh has 8 x 8 cells, 16th mesh has 16 x 16 cells).

The World Wide Merged Cloud Analysis provides total and layer cloud coverage information on a 1/16 mesh (24km resolution), polar stereographic map projection (1440 x 721grid). Cloud cover is calculated based on the percentage of cloudy satellite pixels in each 1/16 mesh grid-box and includes the following: total cloud amount; number of cloud layers (up to four); layer cloud amount; and cloud type, base height, and top height for each layer.

1. Cloud Discrimination

The World Wide Merged Cloud Analysis employs three cloud discrimination algorithms. Each algorithm is unique to the data received from the visible and infrared sensors (OLS) of the Defense Meteorological Satellite Program (DMSP), the Advanced Very High Resolution Radiometer (AVHRR) of the National Oceanic and Atmospheric Administration (NOAA), and geostationary satellite sensors. The algorithms operate in two primary modes, cloud detection and clear-column detection.

The cloud discrimination algorithm for OLS selects a completely cloud-free and a completely cloud-filled pixel to serve as reference values for the subsequent regional cloud cover analysis. The algorithm compares the surface temperature, clear-column infrared brightness temperature, to IR and visible imagery. A surface-temperature model (SFCTMP) provides reference temperatures used in surface-to-imagery comparisons. These temperatures are compared to a database containing expected values given the time of day, location and satellite type. Two thresholds are used to define completely cloud-free and cloud-filled pixels within the analysis region. During the day, pixels below the IR brightness temperature threshold and above the visible brightness threshold are considered cloud-filled. All points between the cloud-free and cloud-filled thresholds are also considered cloud-filled, producing analyses that tend toward cloudy conditions when small or transparent clouds are present.

The multispectral capability of AVHRR reduces the need for accurate background/surface reference information for cloud detection. Snow and sun glint tests are used to minimize the risk of erroneously characterizing reflected surface solar radiation (from sun glint, snow, or ice) as cloud. Snow and ice are identified by comparing a visible channel (0.6 μm) to a channel sensitive to reflected solar and emitted IR (3.7 μm). The latter is also used to identify sun glint. Because it is sensitive to solar reflected radiation, sensor saturation at 3.7 μm normally suggests sun glint. This channel is also used to distinguish between clouds and backgrounds with similar radiative properties.

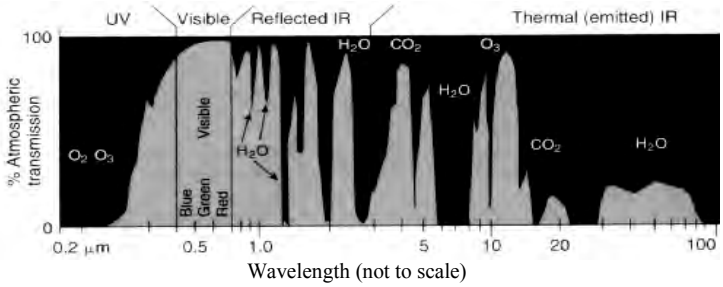


Figure 2. Diagram of atmospheric windows. Chemical notation (CO_2, O_3) indicates the gas responsible for blocking sunlight at a particular wavelength. (After NASA 2011)

Fog and low clouds are detected by differencing the 3.7 μm and 10.5 μm (IR) channels. Liquid water clouds emit and reflect energy at 3.7 μm but are nearly a blackbody at 10.5 μm. Thus, cloud brightness temperatures at night are lower at 3.7 μm than 10.5 μm. The opposite is true during the day. A similar method is used to detect broken and transmissive high clouds at night using 3.7 μm and 11 μm.

Visible ratios are also used in addition to differencing techniques. Clouds reflect similarly at 0.6 μm and 0.9 μm, but backgrounds can vary significantly at these wavelengths. Aerosol scattering can increase reflectance at 0.6 μm (over water) and vegetation can increase reflectance at 0.9 μm (over land). These differences can vary with sun angle and scene composition and in some cases be very large. Because the cloud signal ratio at these two wavelengths is expected to be approximately 1, the algorithm uses a ratio test to distinguish between aerosol/vegetation and cloud signatures.

IR difference tests at 11 and 12 μm are performed to detect ice cloud and small particles along cloud edges. Intrinsic brightness temperature differences in these channels result primarily from dissimilar water-vapor absorption. The presence of ice enhances these differences—ice has a greater extinction at 12 μm. Differences can also occur when the droplet or particle size is smaller than the wavelength. The algorithm compares the difference between 11 and 12 μm to a theoretically derived look-up table to identify ice cloud and small particles along cloud edges.

Cloud detection using geostationary satellite data is performed using temporal differences, cloud-filled thresholds, and spectral tests. The algorithm accounts for cloud movement within the scene by comparing temporal differences or rapid changes in the infrared brightness temperature or visible reflectance. The radiative characteristics of these cloud-filled pixels are then used as a threshold for detecting nearby clouds. Clouds that may have been missed by the temporal difference and threshold tests are detected using spectral techniques similar to those described for OLS and AVHRR.

2. Cloud Layering

Commonly applied cloud layering algorithms for geostationary and polar-orbiter satellite data are used to vertically stratify cloud data in WWMCA. The data are first converted to a 1/16 mesh grid, then a clustering algorithm is used to stratify the pixels into local cloud layers. A cloud type is assigned to each layer using a Real-Time Nephanalysis (Kiess and Cox 1988) technique for cloud typing based on altitude and IR/visible channel spatial variance. The analysis layers are treated as "floating layers," which are allowed to vary in time and space. This reduces discontinuities in cloud layers between adjacent grid-boxes and allows the layers to change mean heights over long distances when satellite images are merged.

The World Wide Merged Cloud Analysis is produced hourly for each grid point on the globe, but the availability of polar orbiting satellite imagery is spatially discontinuous with passes typically occurring 90–20 minutes apart. At grid points where new satellite information is not available, the previous analysis and current imagery must be merged. The initial step in image integration is to compare the age of the analysis to

the age of new satellite information. If the age of new satellite data is more current than any grid cell of the analysis, the new satellite data is incorporated into the analysis. The previous analysis persists in the absence of newer satellite data. When multiple satellite images are newer than a given analysis grid cell, respective satellite image layers are merged and integrated into WWMCA. Persistence of old data results in a lag in cloud advection, development, and destruction compared to reality.

A distance metric is used to distinguish between cloud layers of different satellite images and determine whether cloud layers should be merged or considered distinct. The distance metric is described as follows:

$$D = D_O + \frac{\left(\overline{CTH} - Base\right)}{\alpha_r} \qquad 18$$

where D_o is the base value for D at a height defined at *Base*, from which all other levels are derived. *Base* is the vertical level where D_o is valid, $\overline{CTH}$ is the average cloud top height (meters) of the two layers being compared, and α_r is the response factor, which defines the variance of D away from the base height of *Base*. α_r has the effect of adjusting the height dependence on D, such that as $\alpha_r \to \infty$, then $D\left(\overline{CTH}\right) \to D_O$ (HQ AFWA, 2010).

The distance metric marks the cloud top separation between two layers, which is required for the layers to be considered independent (a non-merge layer). Figure 3 illustrates the distance metric merger technique of two image collections taken by different satellite platforms, Constellations 1 and 2. The integration algorithm selects the satellite image that is the most recent and contains a non-zero cloud amount as the master and labels all others as slaves. A single layer is defined when the separation between the two cloud top heights is less than the distance metric. The top of the new layer becomes the top of the highest cloud layer and the bottom becomes the bottom of the lowest cloud layer.

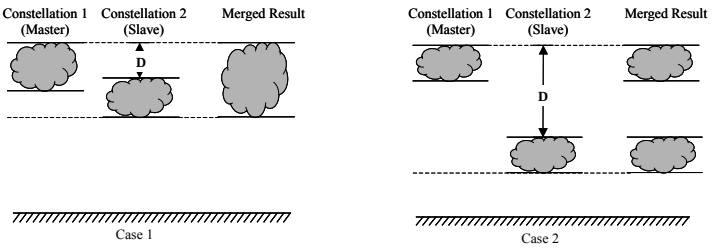


Figure 3. Satellite image merger example. Constellations 1 and 2 are satellite cloud cover data. The most recent data is label master, all others are labeled slaves, and D is the distance metric. (From HQ AFWA 2010)

In cases were more than four cloud layers exist, the layers with the minimum CTH distance are combined as in Case 1 (Figure 4).

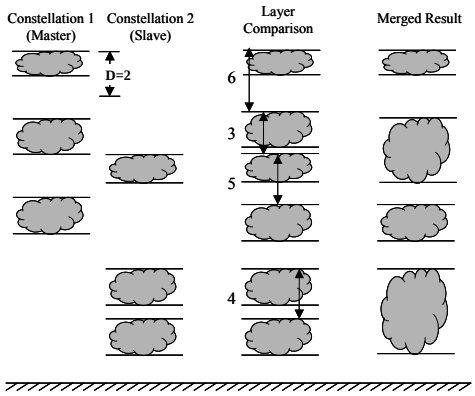


Figure 4. Multi-layer satellite image merger example. Constellations 1 and 2 are satellite cloud cover data. The most recent data is label master, all others are labeled slaves, and D is the distance metric. All satellite retrieved layers are merged into four WWMCA levels. (From HQ AFWA 2010)

B. THE SHORT-RANGE CLOUD FORECAST

The Short-Range Cloud Forecast (ADVCLD) is the forecast segment of CDFSII. Global cloud forecasts are produced on a polar-stereographic grid centered on 80°W in the Northern Hemisphere and 100°E in the Southern Hemisphere. Forecasts are typically produced from 6 to 30 h in 3-h increments for five vertical levels: gradient level (60 mb above surface), 850, 700, 500, and 300 mb. Floating cloud layer information from

WWMCA (cloud amount, base and height) is assigned to fixed layers in ADVCLD and used as the initial cloud condition. The final output of the model includes total cloud cover to the nearest one percent.

Figure 5 illustrates the conversion of the four WWMCA floating layers to the five ADVLCD fixed layers. The standard atmospheric heights are assigned to pressure levels and compared to layer tops and bottoms. Above ground cloud layers range from surface to 700 mb, surface to 500 mb, and surface to 300 mb. When a WWMCA cloud value intersects one of the standard layers of ADVCLD, ADVCLD adopts the WWMCA value at that level. If two or more WWMCA cloud layers intersect one of the standard layers of ADVCLD, the WWMCA layer containing the maximum cloud fraction is used for the conversion. Once all WWMCA layers are converted, the ADVCLD layer with the maximum cloud fraction is used as the total cloud amount of the grid point.

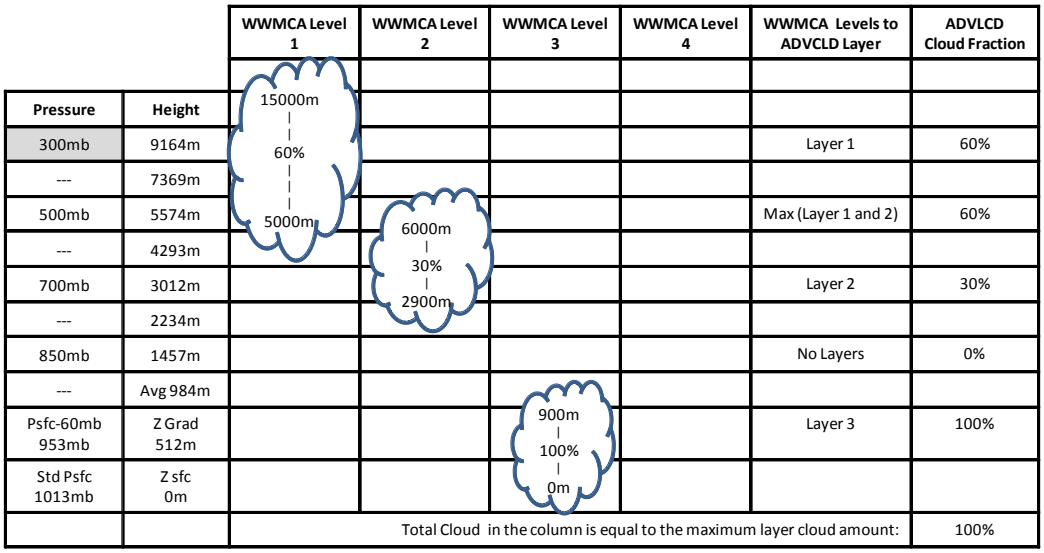


Figure 5. WWMCA to ADVCLD layer conversion. The four WWMCA levels are converted to five ADVCLD layers. The layer with the maximum cloud fraction value is used to define the total cloud amount. (After McDonald Personal Communication).

The model uses a Quasi-Lagrangian advection technique. Horizontal cloud trajectories are calculated using u and v wind components from a global numerical weather prediction (NWP) model. Vertical cloud trajectories (parcel displacement from

one pressure level to another) are converted from global NWP vertical pressure velocity (omega) values. Horizontal and vertical trajectories are used as advection mechanisms.

Cloud amounts from WWMCA are not used directly in the advection scheme. Condensation pressure spread (CPS) is the prognostic variable. It represents the pressure difference between the cloud fraction in a particular grid box and its lifting condensation level. Values of CPS near zero are considered 100% cloudy, and large CPS values are considered clear. The CPS values are calculated from NWP dew point depression values and compared to WWMCA CPS values at each grid point and the largest (driest) of the two is chosen. Therefore, initialization values can be slightly different from WWMCA values.

Vertical motion produces variations in the CPS distribution, which produces variations in layer cloud amounts. For example, advection into a ridge decreases CPS values and subsequently increases the cloud fraction within the grid box. Advection into a trough increases CPS values resulting in a drying affect. Cloud to moisture curves first derived in an empirical study by Essenwanger and Haggard (1961) are used to convert CPS values to cloud fractions. The current values, modified by AFWA, are plotted in Figure 6. Condensation pressure spread values are compared to values contained in a look-up table and converted to cloud percentages at respective pressure levels (850 mb, 700 mb, 500 mb, and 300 mb).

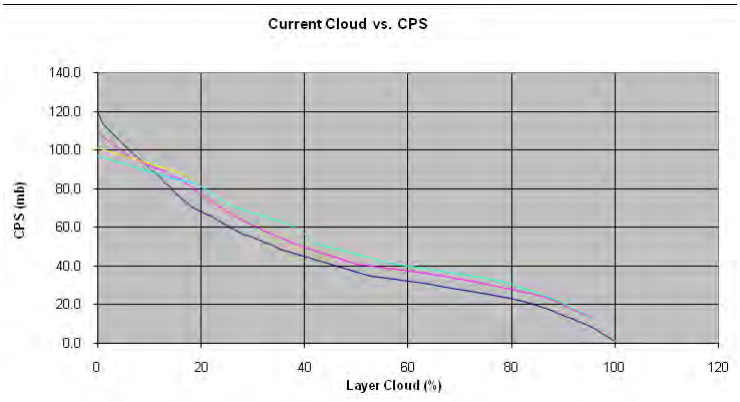


Figure 6. Cloud conversion table is used to convert 300 mb (blue), 500 mb (yellow), 700 mb (violet), and 850 mb (dark blue) condensation pressure spread values to cloud fractions (%). (From HQ AFWA, 2010).

C. GLOBAL ENSEMBLE FORECAST SYSTEM

The Global Forecast System (GFS), produced by the National Centers for Environmental Prediction (NCEP), is AFWA's model of choice for NWP parameters within ADVCLD; therefore, we have chosen to use the Global Ensemble Forecast System (GEFS) to drive the Global Advection Cloud Ensemble (GACE). GEFS uses an ensemble transform method (ET) with rescaling (ETR) to define the initial atmospheric uncertainty (Wei et al. 2008). Adapted from the ET method devised by Bishop and Toth (1999), the background error covariance is used with the short-range forecast spread to produce perturbations and the error variance of the analysis from the NCEP data analysis system is used to scale the magnitude of the initial perturbations. Using the error variance of the analysis, however, does not sufficiently limit the initial perturbations at extended lead-times. Therefore, regional rescaling is applied to the error variances of each grid point to further restrain the initial ensemble spread. The ETR method replaced the breeding method in GEFS during NCEP's May 2006 implementation.

In February 2010, NCEP made additional adjustments to the methods used in developing initial ensemble perturbations. The most significant was the introduction of a stochastic total tendency perturbation scheme (STTPS). This approach, originally proposed by Buizza (1999), attempts to account for random model errors by imposing stochastic terms on the total tendency equations (Hou et al. 2010). The primary goal for implementing STTPS was to improve the forecast by increasing the ensemble spread, yet reducing the number of outliers.

In the fall of FY2010, NCEP made additional changes to the ensemble configuration. Model initialization was converted from GFS V8.00 to V9.01. This was done to improve the ETR initialization and the stochastic total tendency perturbations. GFS V9.01 includes adjustments to moisture algorithms and background error computations. We suspect that this implementation has implications on the ensemble spread of the latter months of our dataset.

Currently, NCEP runs GEFS four times per day, 0000 UTC, 00600 UTC, 1200 UTC and 1800 UTC. The system produces 20 perturbed and one control forecast per

cycle with a lead-time of 16 days. The ensemble forecasts are made available at 1 degree resolution and 6 h time-steps. Considerations are being given to increasing horizontal resolution to 0.5 degree and temporal resolution to 3 h in the fall of 2012 (Wobus 2011 and Zhu 2011).

The Receiver Operating Characteristic (ROC) area curves in Figure 7 show the NH 10 meter u and v winds, which result from pre- and post-resolution changes. The ensemble demonstrates most skill within the first 24 h of the forecast period. Although the skill of the higher resolution ensemble (red) is degraded in the first 4–5 days, the increased resolution between time-steps could improve ADVCLD trajectories and subsequent cloud advections.

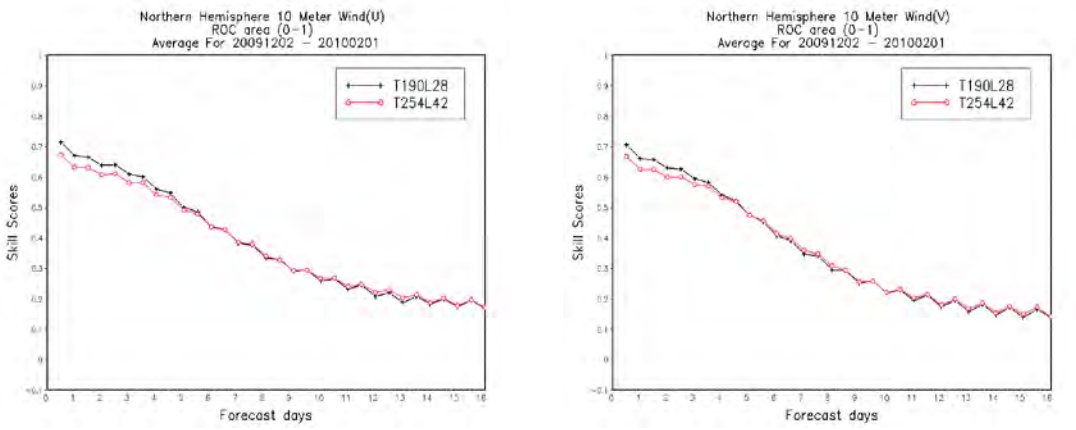


Figure 7. GEFS skill scores for 10 meter u (left) and v (right) winds. Depreciation in skill is expected with the .5 degree resolution ensemble (red) as compared to the 1 degree resolution ensemble (black) within the first five days of the forecast period. (From Zhu et al. 2011)

GEFS is not designed for, nor has it been tested for, cloud-free forecast applications. This offers a unique challenge in that the ensemble perturbations are geared towards optimizing uncertainty characterizations to improve forecasts at extended lead-times. Uncertainty in cloud cover, however, can become significant very quickly. In the presence of convection, clouds can develop and decay on the order of hours not days.

D. GLOBAL ADVECTION CLOUD ENSEMBLE

We combine WWMCA and GEFS using the ADVCLD initialization and advection algorithms to produce a Global Advection Cloud Ensemble (GACE). Although ADVLCD is produced using GFS data at half degree with a 3-h time-step, GEFS limits our resolution to one degree with a 6 h time-step. In addition, the WWMCA analysis is interpolated to one degree resolution to match GEFS' spatial resolutions during the forecast initialization of the ensemble control forecast. The lack of complex physics in cloud development, the simple advection scheme, and ensemble initialization limits the use of the GACE to large scale motions and features.

Here we show that GACE is temporally and spatially consistent with WWMCA. Large differences in model and analysis cloud amounts over time indicate that model biases exist relative to the analysis. The global cloud amounts in our system should remain consistent throughout the forecast cycle because GACE is an advection model lacking complex cloud production or dissipation physics. Never-the-less, we compare ensemble variations in the cloud cover mean, standard deviation, and autocorrelation to those found in the analysis. This will provide a first look into the ability of GACE to predict the cloud evolution of the atmosphere.

1. Mean Cloud Cover

The global mean cloud cover is calculated every 6 h for the ensemble control, ensemble mean, and analysis (WWMCA) for the 0000 UTC and 1200 UTC cycles. Forecasts are sampled every three days and the analysis is sampled daily. At each forecast hour (0, 6, 12, 18, and 24), the daily mean cloud-cover of all grid points is calculated, and then the monthly mean cloud-cover is computed. The analysis mean is also used for the 24-h analysis mean. These steps are repeated for all months in the data set.

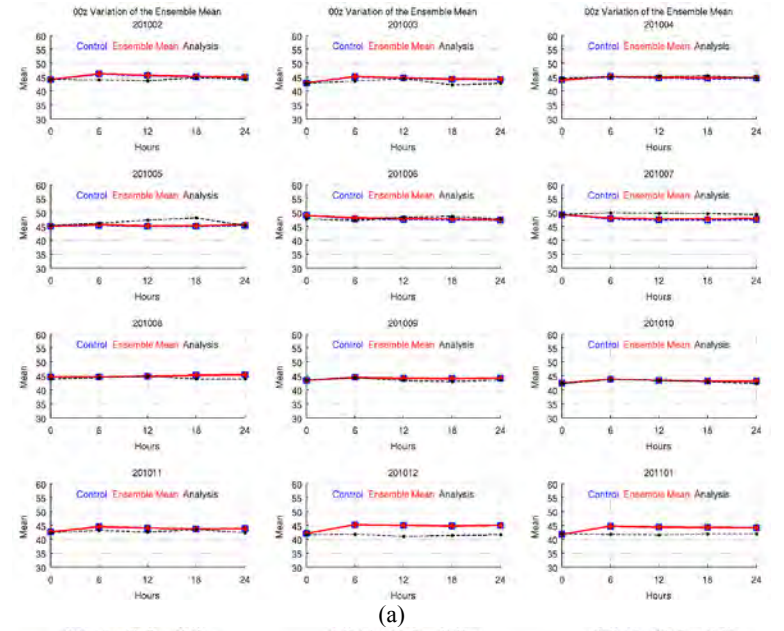
Figure 8 contains monthly plots of the hourly variations in the mean cloud-cover across the globe. Large temporal variations make the mean unrepresentative of future states of the atmosphere. If the mean increases (decreases) with time, the sample mean grows (shrinks) with respect to sample size, and the forecast will always underestimate

(overestimate) the mean at some time in the future. The monthly mean cloud cover for the ensemble mean forecast, control forecast, and analysis remain nearly stationary at about 45% $\pm$ 5% throughout the 24-h forecast period for both the 0000 UTC and 1200 UTC forecast cycles.

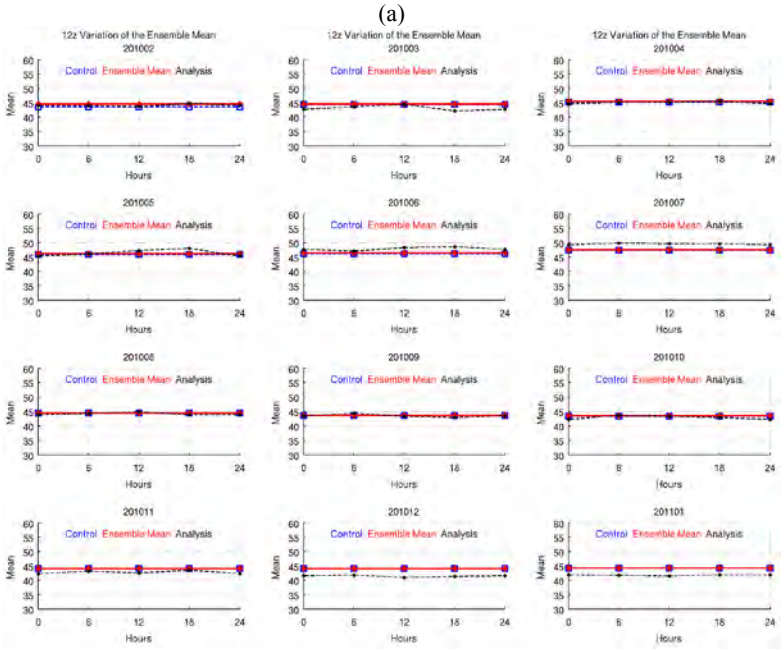
Figure 8 also indicates that the ensemble has seasonal moist and dry biases. The ensemble 6-h forecast, and beyond, demonstrates a slight moist bias for the 0000 UTC cycle during the months of February and March. The February bias, however, is not apparent with the 1200 UTC cycle. During the months of May, June, and July, the ensemble exhibits a dry bias for both 0000 UTC and 1200 UTC cycles, but the 1200 UTC cycle bias is less pronounced. No significant bias is noted during August to October, but the moist bias reemerges from November to January.

These seasonal trends in the bias coincide with winter and summer fluctuations of the position and zonal flow patterns polar front jet. Biases are generally negative during the summer when flow is largely zonal and positive in the winter when flow becomes more meridional. These biases combined with the design of the ensemble suggest that the ensemble has a natural dry bias but over forecasts clouds in the presence of synoptically induced vertical motion.

The tendency for the ensemble to over or under forecast cloud cover as described above can be real or a function of sample size. Our comparisons are based on data sets that have different sample sizes, which have implications on the magnitude of the mean. In addition, the mean cloud cover can reasonable vary with region. Therefore, we further evaluate the ensemble biases identified here when we perform our regional analysis of skill in Chapter V.



(a)



(b)

Figure 8. Time-series of (a) 0000 UTC cycle and (b)1200 UTC cycle variation in the global mean cloud cover of the ensemble mean (red dot), ensemble control (blue square), and WWMCA (dashed line) from February 2010–January 2011. Winter moist bias and summer dry bias in ensemble mean and control forecasts.

2. Spatial Variance

The spread about the mean cloud cover can be measured by the standard deviation, which is the square root of the variance. The standard deviation represents the normal displacement from the mean value. Therefore, values outside the "norm" are considered large or small. Changes in the standard deviation, as with the mean, cause problems in data substitutions and representations. If the standard deviation changes with time, values currently within (outside) the normal displacement of the mean may fall outside (inside) the norm at some future time.

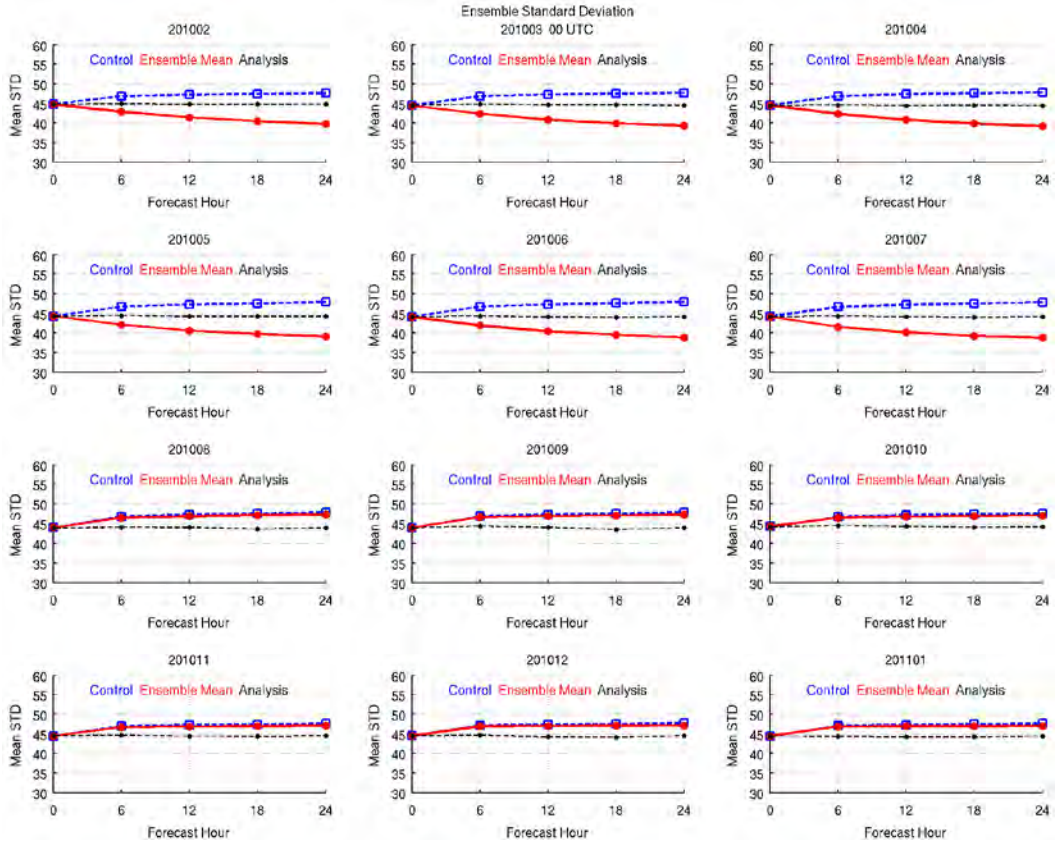


Figure 9. 0000 UTC time-series of global spatial cloud variation (STD). Ensemble mean (red dot) decreases with time, ensemble control (blue square) stationary about 47%, and WWMCA (dashed line) stationary about 45% from February 2010–January 2011.

Figure 9 shows the monthly standard deviation of global cloud cover from the control forecast, mean forecast and analysis (WWMCA) for the 0000 UTC cycle. To calculate the standard deviation, we begin with the variance. First, we calculate an hourly variance across the globe for each day (every three days with the forecasts). Next, we calculate the mean of the daily variances. Finally, we take the square root of the mean variance to calculate the standard deviation. The process is repeated for each 6 h time-step through 24 h and each month in the dataset.

Significant changes do not exist in the standard deviation for each forecast hour of the 0000 UTC or 1200 UTC cycles (not shown). The standard deviation of the analysis is stationary at about 45%. Thus, cloud-cover values are generally found within $\pm$ 45% of the mean. The standard deviation of the control forecast tends to be higher than the analysis and mean forecast. As we will see in section D.4 of this chapter, WWMCA has more mid-range (10%–90%) cloud-cover observations than the ensemble forecast. The mid-range values decrease the standard deviation as compared to the control forecast, which loses the majority of these mid-range values at the first time-step and tends towards 0% and 100%.

The standard deviation of the ensemble mean forecast equals the analysis at time 0 but quickly decreases with forecast lead-time. This represents the convergence of the forecast towards the mean value, which is driven primarily by forecasts preferring 0% and 100% cloud fractions. This suggests that clear or cloudy preferences become stronger with lead-time. The ensemble mean performs similar to the control during the latter 6 months.

The abrupt change in the ensemble is coincident with the NCEP implementation accomplished in July 2010. Figure 10 is a plot of the mean global spread of the ensemble members for each month of our dataset. The variance between ensemble members is taken at each grid point, and the mean variance is calculated. Our plot shows the global standard deviation (square root of the variance). Although the global spread of the ensemble members is substantially small as demonstrated with the ensemble mean, the spread drops to zero after August. Therefore, we point to the NCEP implementation as a possible catalyst for the lack of diversity in the ensemble members after August.

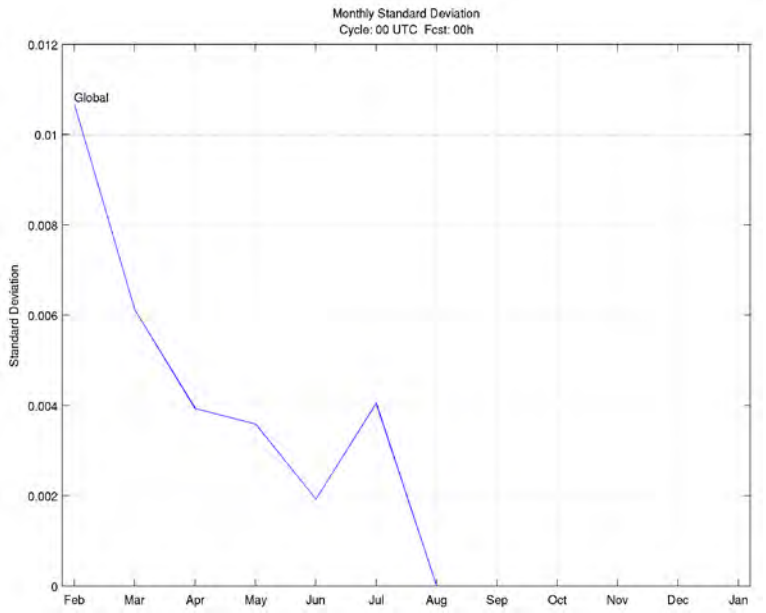


Figure 10. Variations in the monthly standard deviation at the initial forecast hour are calculated to evaluate to evaluate the changes in the ensemble spread. Spread decreases to zero in August.

We learn four principle things from our spatial variability evaluations. First, the analysis is stationary at about 45% cloud cover variability. Second, the tendency to forecast extreme values increases the standard deviation of the control over the analysis—because cloud cover percentages are rounded to 0% and 100%. Thirdly, the ensemble mean forecast may have seasonal cloudy or clear preference. Forth, the significant reduction in the ensemble spread coincides with an implementation performed by NCEP during July 2010. The key point, however, is the standard deviation of the global cloud cover analysis is sufficiently stable for cloud-free forecast verification.

To get a better feel for atmospheric variability, it is useful to calculate the hourly (Figure 11a) and monthly (Figure 11b) standard deviations. We can also see from these charts that the analysis is consistent, but a downward trend in the analysis can be seen as we move into the month of March. As a result, model evaluations may need to be binned by seasons or even months before averaging to ensure forecast assessments are properly represented.

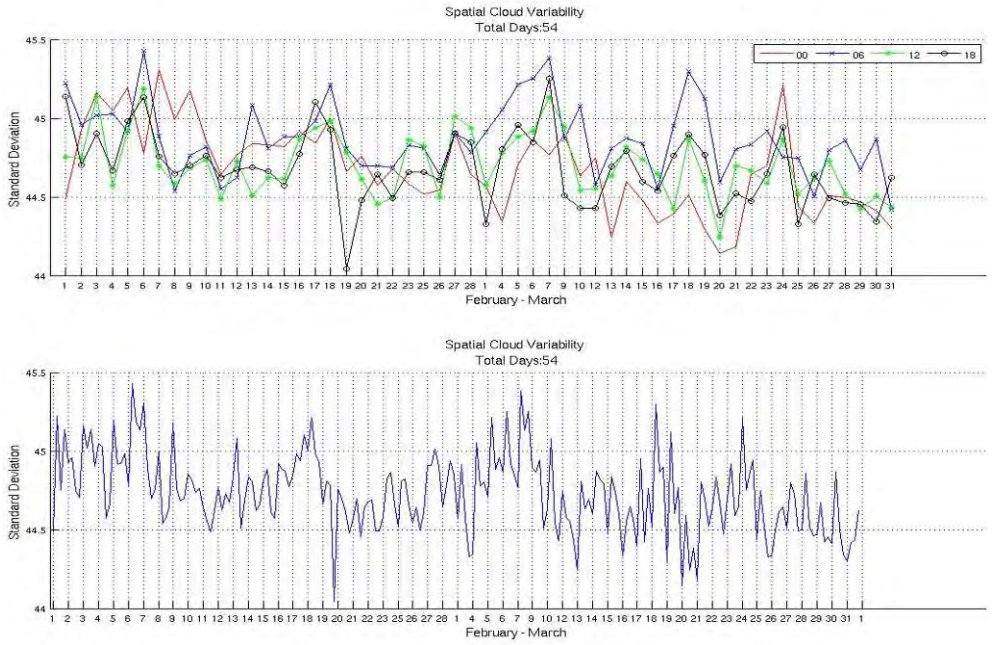


Figure 11. Time-series of spatial cloud variation during the month of February and March 2010. Data is plotted in six hourly increments. Hourly data is separated (top) and combined (bottom).

3. Autocorrelation

Knowing GACE and WWMCA are independent engenders confidence that WWMCA can be used for verification. Autocorrelations provide insight about predictability time scales and temporal independence, and they can be calculated using the lag-*k* autocorrelation coefficient function.

$$r_k = C_k / C_s \tag{19}$$

$$C_k = \frac{1}{N}\sum_{i=1}^{N-k}\left(x_i - \bar{x}_-\right)\left(x_{i+1} - \bar{x}_+\right) \tag{20}$$

$$C_s = \frac{1}{N}\left[\sum_{i=1}^{N-k}(x_i - \bar{x}_-)^2 \sum_{i=1}^{N-1}(x_{i+1} - \bar{x}_+)^2\right] \tag{21}$$

where C_k is the auto-covariance coefficient (20) and C_s is the product of the standard deviations (21). The subscripts "–" and "+" represent the first and last N–*k* lag-*k* autocorrelation. This autocorrelation function was computed for each global grid point (1440x721) for k = 0, 6, 12, 18, and 24.

The time-series at Figure 12 shows little-to-no correlation at *k*=6 for the control/ensemble mean forecast and cloud analysis. Essentially, this indicates that the two datasets have independent cloud cover distributions. Verification of GACE's 6-h forecast against the 0600 UTC analysis will yield results with minimal dependence (in an Eulerian frame of reference) on the initial forecast state (0000 UTC analysis).

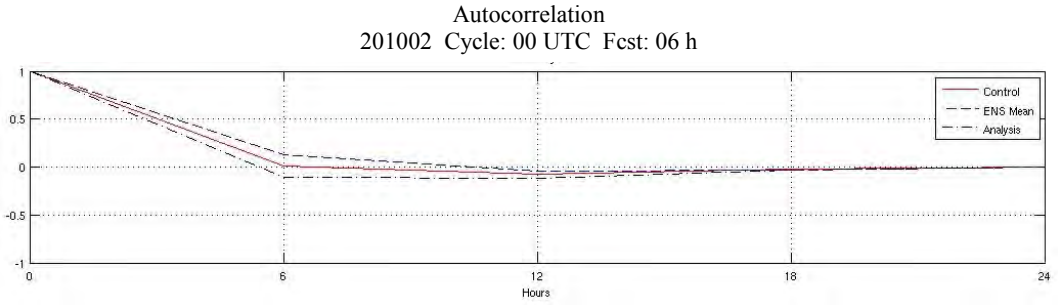


Figure 12. Time-series of 0000 UTC cycle variations in the global autocorrelation function in cloud cover during the month of February. 1200 UTC cycle not shown due to lack of significant difference from 0000 UTC cycle.

4. Dispersion

Although ensembles are useful in capturing uncertainty, they are derived from an imperfect model and are subject to systematic errors. Errors in model initial conditions, physics, discretization, analysis error, and lateral boundary conditions are sources of systematic error. Therefore, beginning with an imperfect model and perturbing the model with less than optimal initial and lateral boundary conditions will most likely produce an uncalibrated ensemble (Raftery et al. 2003), where ensemble forecast distribution is not consistent with climatology. Manifestations of an uncalibrated ensemble are model biases (e.g., consistently over/under forecasting cloud cover) and inaccurate error growth (under-/over-dispersion).

Hamill (2000) demonstrated how verification rank histograms (VRH) can be used to test how well uncertainty is represented by an ensemble. If an ensemble is based on initial perturbations that are equally likely, then the resulting forecasts should be equally likely. Furthermore, observations that are plausible members of the ensemble forecast

indicate that the ensemble is reliable. In a perfect or well-calibrated ensemble, all members share a similar likelihood of occurrence which results in a uniform verification distribution.

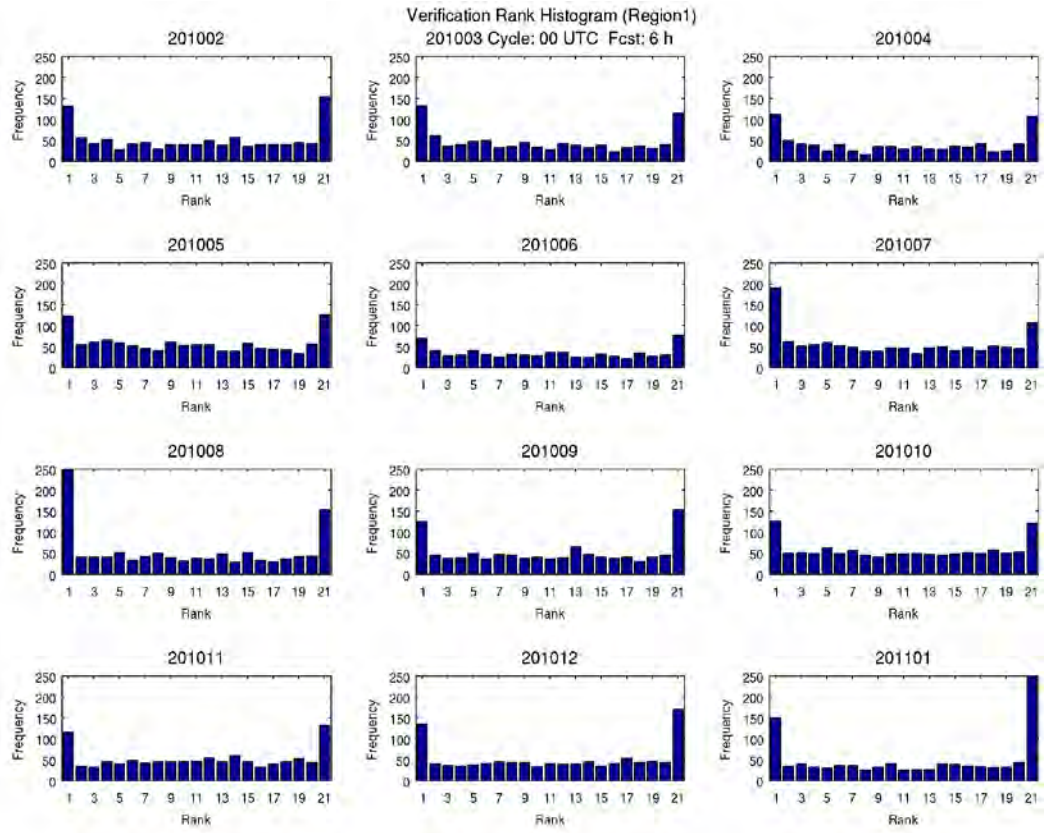


Figure 13. Verification rank histogram of 20 member ensemble cloud-free forecasts for Region 1.

Verification rank histograms are constructed with member (N) + 1 bins. Each ensemble forecast and analysis value is sorted from lowest to highest and a rank is assigned to the bin in which the analysis falls. The lack of unique values in the cloud fraction forecasts among ensemble members required the introduction of random deviates (Press et al. 1992). At set of N+1 uniformly distributed pseudorandom numbers are generated and added to N+1 forecasts. The random deviates are significantly small random uniform numbers (.0000–.0100) used to create delineation between like forecasts and to avoid impacting the bin assignment of values differing from the analysis.

Figure 13 shows the monthly rank histograms for the 0000 UTC cycle and 6 h ensemble forecast. The verification rank histograms are not global, but taken from one of the regions we later select for forecast evaluations. As expected, our VRHs are not uniform. The most populated bins tend toward the extremes. This is due in large part to the fact that modeled, as does the observed, cloud cover tends toward 0 and 100% (Figure 14). Comparing Figure 14a to Figure 14b, distributions of clear vs. cloudy conditions between WWMCA and GACE are very similar with respect to the extreme values of 0% and 100% cloud cover during the month of May—other months yield similar results.

The model rarely forecasts cloud fractions other than clear and cloudy conditions which we discuss further in the method section. Deviations from the extreme cloud cover values seen in the analysis (Figure 14a) represent the edges of large scale cloud features rather than distinct clouds themselves. These values are most often considered clear by the ensemble.

We have ensured adequate distance exists between sampled points and therefore expect the forecast solutions to be reasonably independent (Hamill 2000). Variability among ensemble members depends on the spatial and temporal correlation of errors in the model. Furthermore, non-uniform rank distribution can arise from errors introduced during the initialization of the system, bias in the model moisture or wind fields, and/or errors in the verification analysis.

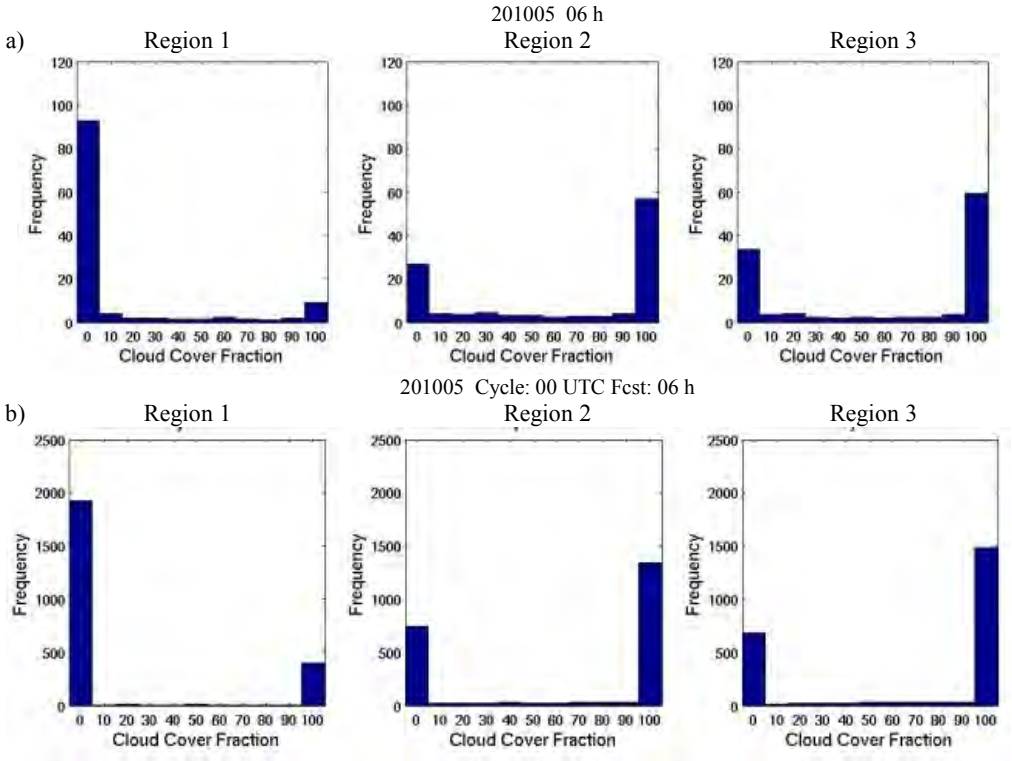


Figure 14. Frequency of cloud cover. The frequency of the cloud cover in a) WWMCA and b) ensemble are comparable. The values in each bin represent the cumulative cloud cover fraction between each interval by region for the 6-h forecasts from 0000 UTC cycle (except extreme bins). Cloud fractions 0 and 100 are counted independently.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. METHODS

A. SELECTING REGIONS

The three shaded regions in Figure 15 are the focal points of our research. These regions were selected based on cloud-cover frequency, operational significance, and latitude. Region 1 covers Saudi Arabia and Iran and is most frequently clear. Region 2 covers China and has variable cloud-cover conditions. Region 3 covers the northern portion of South America and is most frequently cloudy. Each region has operational significance and lies between 60N and 60S, which avoids the challenges of foreshortening due to satellite geometry.

Conditions are defined as clear when the model or observed cloud fraction is less than or equal to 30% and cloudy when the cloud fraction is greater than 30%. We use two steps in establishing these regions with respect to cloud-cover frequency. First, we perform a broad scale evaluation of cloud cover across the globe to choose regional candidates. Then, we use local calculations to refine our selections.

(a)Region 1 (b) Region 2 (c) Region 3

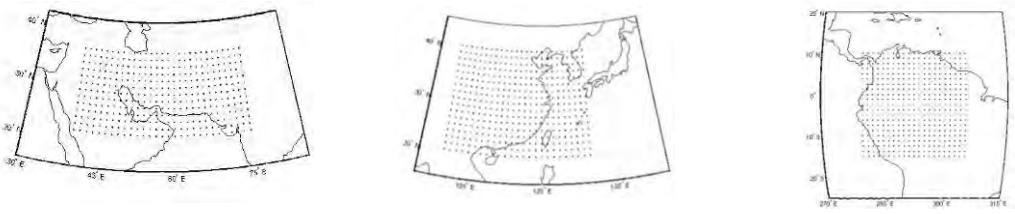


Figure 15. Shaded areas represent the regions selected for forecast analysis based on annual frequency of cloud cover, operational significance and latitudinal location.

1. Regional Selection

The daily cloud analyses from February 2010 to January 2011 are used to calculate the monthly frequency of cloudy conditions at each grid point across the globe.

$$Fr(M)_{hour} = \frac{1}{N}\sum_{i=1}^{N} M(obs)_i \qquad M(obs) = \begin{cases} 0, & if\ obs \le 30\%, \\ 1, & if\ obs > 30\%; \end{cases} \qquad 22$$

$$hour = 0,6,12,18$$

where Fr(M) is the monthly frequency of cloudy conditions at a given grid point and N is the number of days per month. M(obs) is a step function that jumps from 0 to 1 when the observed cloud cover is equal to 30%.

The mean cloud cover Fr(M) is used to determine whether cloud conditions are most often clear, variable, or cloudy. Months with Fr(M) $\leq$ 30% are considered clear; months with 30% < Fr(M) $\leq$ 70% are considered variable; and months with cloud cover frequencies Fr(M) > 70% are considered cloudy. Clear, variable, and cloudy months are totaled, respectively, to assess the annual persistence of each cloud-cover condition.

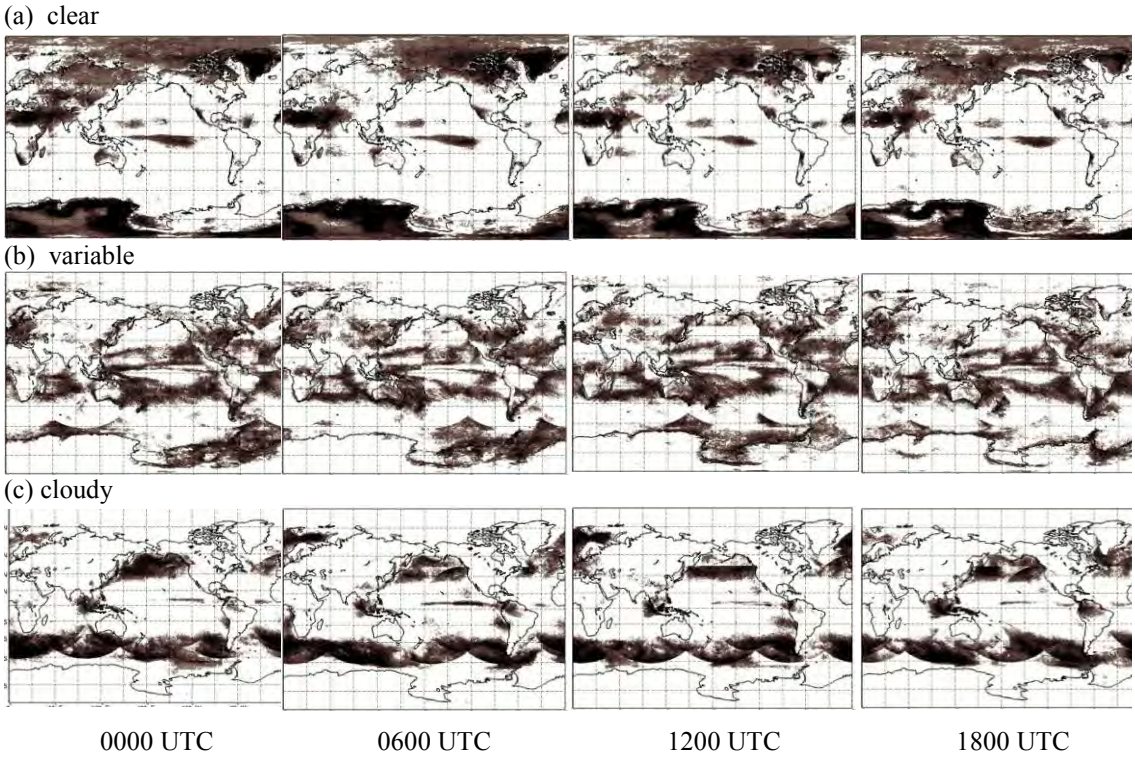


Figure 16. Analysis of cloud cover frequency used to identify prospective cloud cover environments. Shaded areas indicate regions where monthly frequency of clear (a), variable (b), and cloudy (c) conditions occurred for more than 6 months during the 12-month analysis period.

Figure 16 provides a global view of annual cloud cover frequency and a reasonable first guess in the selection of our clear, variable, and cloudy regions. Region 1 is selected from the large clear area that stretches from East Africa to Afghanistan (Figure 16a). This desert area is a good candidate for persistently clear conditions. For a variably cloudy region, we desire a location that favors neither cloudy nor clear conditions. Here we choose the region covering China and North Korea (Figure 16b). North America is much larger and has a stronger variable cloud cover signal, but Region 2 is preferred for its operational significance. Likewise, the North Pacific Ocean and the Maritime Continent are most suitable for examining cloudy conditions (Figure 16c), but South America is elected as our third region (Region 3) for its significance in land operations.

2. Regional Refinement

Each region is confined to an approximate 2400 km area (10,000 grid points) using the following formula.

$$(x_2 - x_1)(y_2 - y_1) = 10{,}000 \tag{23}$$

$$y_2 = \frac{10{,}000}{(x_2 - x_1)} + y_1 \tag{24}$$

The easternmost, westernmost, and northernmost points (x_1, x_2, and y_1) are chosen explicitly, and then the southernmost point (y_2) is calculated. Once the regional boxes are calculated, we separate them into 24 km sub-regions. To limit spatial correlations between the grid points, we choose grid points that are equally spaced at 240 km (10 grid points). The sufficiency of our displacement between grid points is confirmed by the results of an autocorrelation displayed in Figure 17. An autocorrelation is performed on each sub-grid with its surrounding grid points from a distance of 1 to 10 grid points. Northeast, northwest, southeast, and southwest distances are calculated using the Pythagorean Theorem. Figure 17 indicates that cloud cover correlations drop to zero at a distance of 7 grid points in Region 2 and 3. Because clouds are rarely observed in Region 1, the correlation between grid points does not reach zero at our selected distance, but is significantly small.

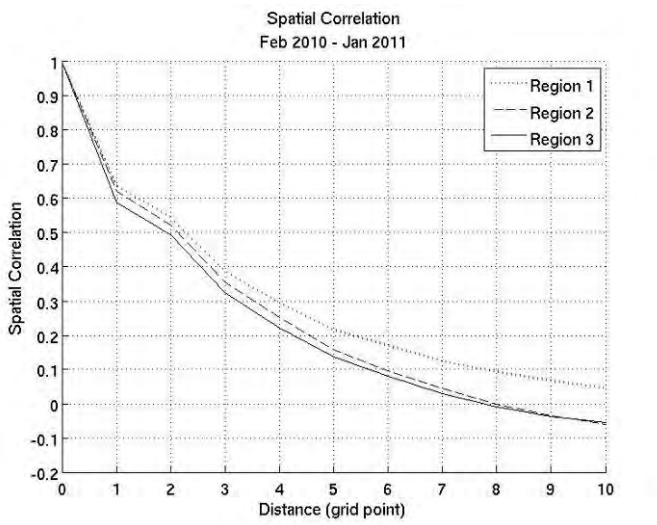


Figure 17. Spatial correlation of cloud cover between regional grid points. Spatial correlations decrease with distance. Each grid point distance equates to 24 km.

Once the regions are defined, we charted the monthly frequency of cloudy conditions over the entire domain (region) to evaluate whether the regions are truly representative of clear, cloudy or variable cloud cover. Figure 18–Figure 20 displays the fraction of the domain affected by the frequency of cloudy conditions within the domain. Figure 18, indicates cloudy conditions less-than 30% in Region 1 on a month-to-month (thin lines) and annual (thick line) basis. In Region 2 (Figure 19), significant peaks are noted outside the variable threshold (30% < Fr > 70%), but the largest area under the curves falls within our definition of variable cloud conditions. Figure 20 clearly indicates that cloudy conditions dominate Region 3.

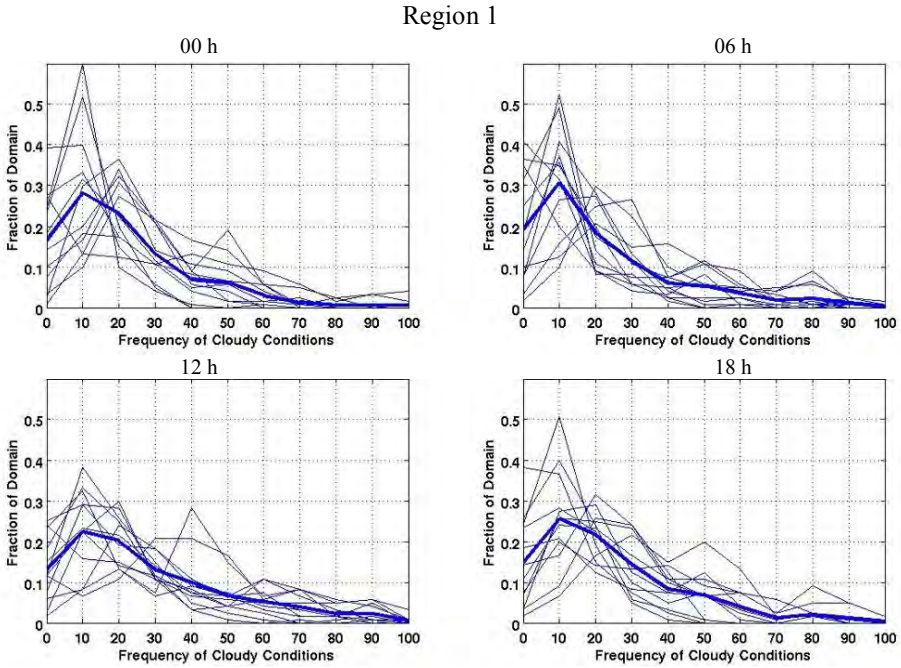


Figure 18. Frequency of cloudy conditions in Region 1. Charts display the fraction of the domain affected by cloudy conditions ranging from 0 to 100% of the time. Thin lines are monthly domain variations, and the thick lines are the annual domain variations.

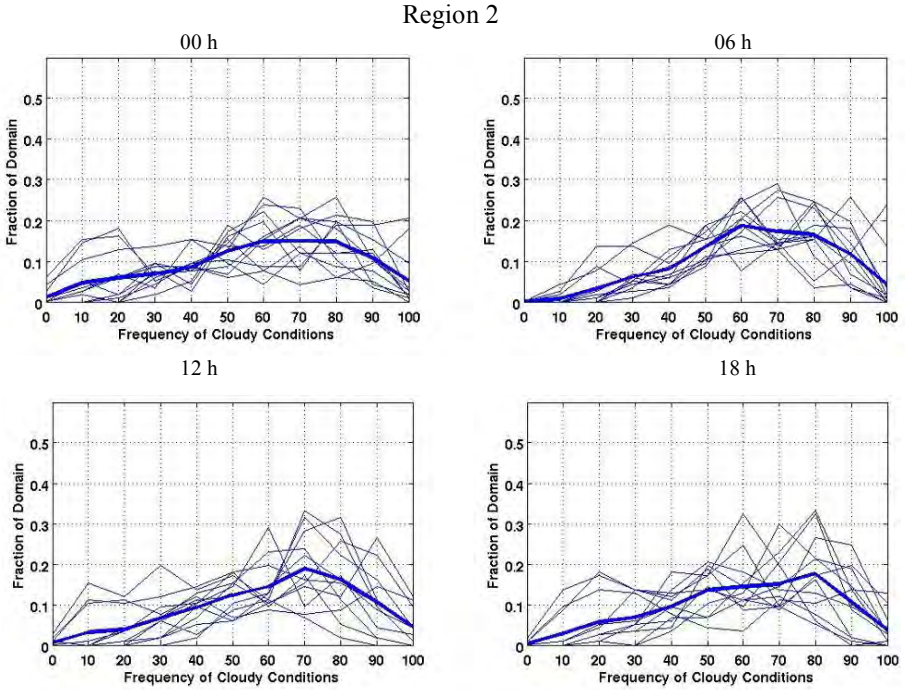


Figure 19. Frequency of cloudy conditions in Region 2. Charts display the fraction of the domain affected by cloudy conditions ranging from 0 to 100% of the time. Thin lines are monthly domain variations, and the thick lines are the annual domain variations.

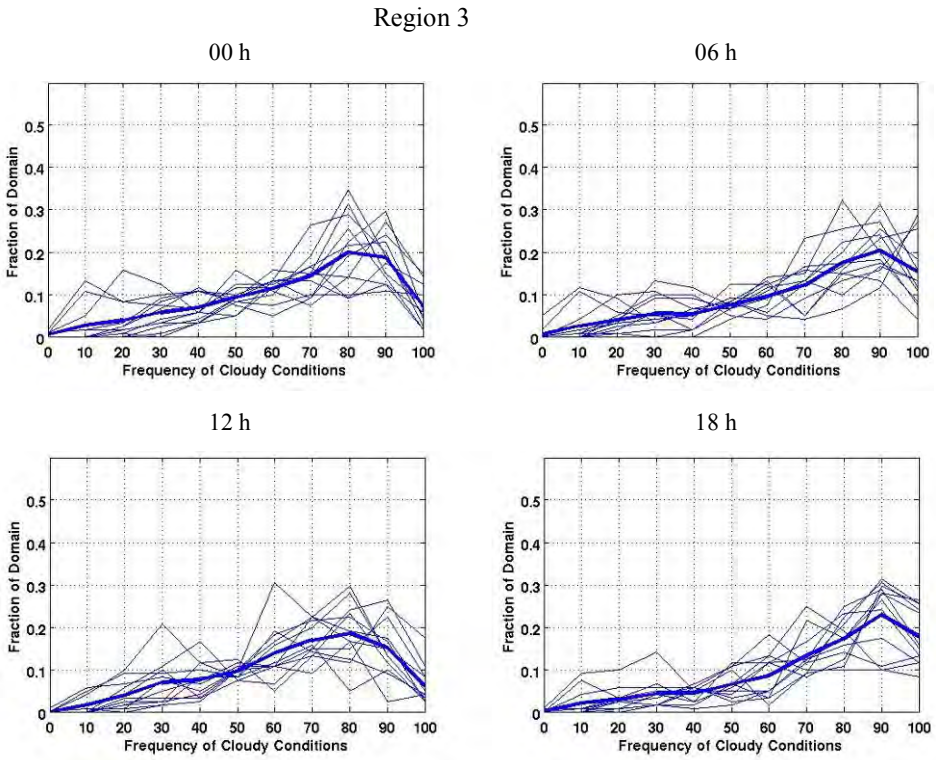


Figure 20. Frequency of cloudy conditions in Region 3. Charts display the fraction of the domain affected by cloudy conditions ranging from 0 to 100% of the time. Thin lines are monthly domain variations, and the thick lines are the annual domain variations.

One of the challenges we face in this analysis is our limited data set. Calculations that are intended to represent climatology are typically performed on data sets that span many years, preferably 30 years or more. In an effort to test the consistency of our results, we perform the two-sample Kolmogorov-Smirnov (K-S) test (Wilks, 2006). This test compares the cumulative frequency distribution of two data sets to determine if they could possibly come from the same population.

Figure 21 shows K-S test results for February 2010 and February 2011. We start with the null hypothesis that 2010 data are from the same population as the 2011 data. We discover that the likelihood of cloud cover being similar for both months is 99% or more in each region. This comparison, although limited to a single month, gives us

confidence that our results may loosely represent the climatological norm of each region if not for years similar to 2010. Now that our regions are established, we turn our attention to the ensemble.

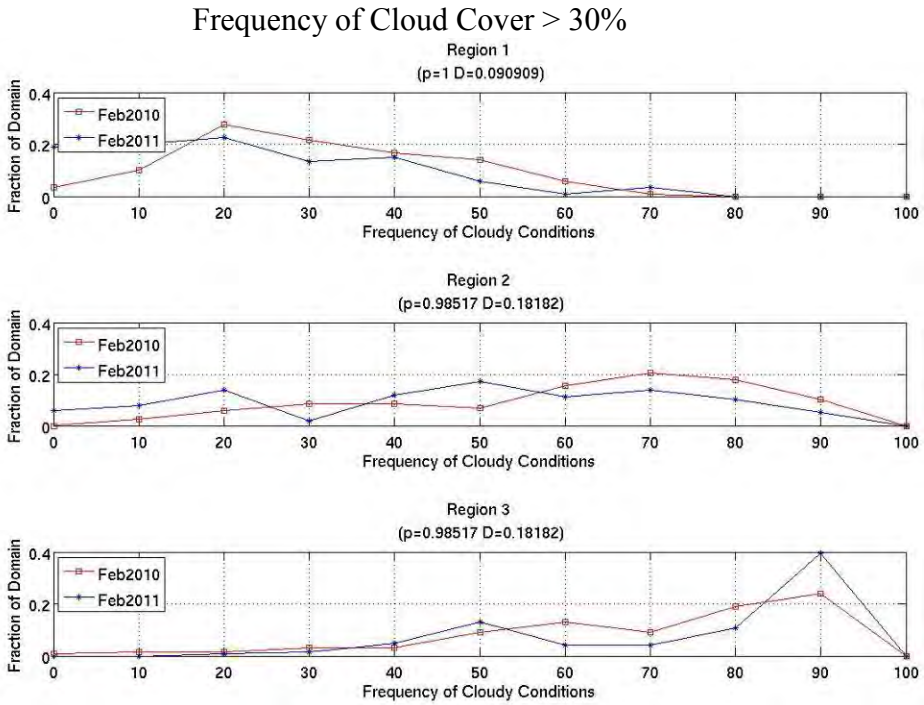


Figure 21. K-S test between February 2010 and February 2011. Frequency of cloud cover across the region is plotted for visual comparisons. Null Hypothesis: No difference exists between the frequency of cloudy conditions within each region. High p-values (max=1) suggest the null hypothesis is true.

B. CALCULATING ENSEMBLE PROBABILITY

We calculate ensemble probability forecasts via three methods: democratic voting (Doran, 1997), uniform ranks (Hamill and Colucci, 1997), and weighted ranks (Eckel, 1998). Here we provide a description of each and a hypothetical example to illustrate their behaviors. Although GACE is a 20-member ensemble, we do not need to consider all 20 members to examine the attributes of these methods. Therefore, we limit our discussion of ensemble probability forecasts calculations to a hypothetical 6-member ensemble with cloud fraction forecast values of 17, 18, 20, 27, 33, and 36.

Democratic voting is the most direct method of calculating ensemble probability forecasts because each member is given equal weight. First, we establish the number of

required bins. Each bin represents a verification threshold. Since we are only interested in cloudy or clear conditions, we only require two bins. Next, the number of member forecasts that fall into each verification bin is tallied. Finally, each bin is divided by the total number of forecasts. Table 4 shows the ensemble probability of clear conditions that would result from our hypothetical scenario. With 4 of the 6 members forecasting clear conditions, the resulting ensemble forecast probability is 67%.

Table 4. Democratic voting method of calculating probability using hypothetical ensemble values.

Bin	Verification Range (Cloud-Fraction)	Number of Ensemble Members in Range	Probability of Calculation
1	≤ 30	4	$P_1 = 4/6 = .67$
2	≥ 30	2	$P_2 = 2/6 = .33$

The uniform ranks method better approximates forecast probability (Hamill and Colucci 1997) because it accounts for the probability that observations will fall between ensemble-member forecasts. This method assumes each ensemble member (n) is equally likely to verify. Therefore, the verification is equally likely to fall into any n+1 bin (Figure 22).

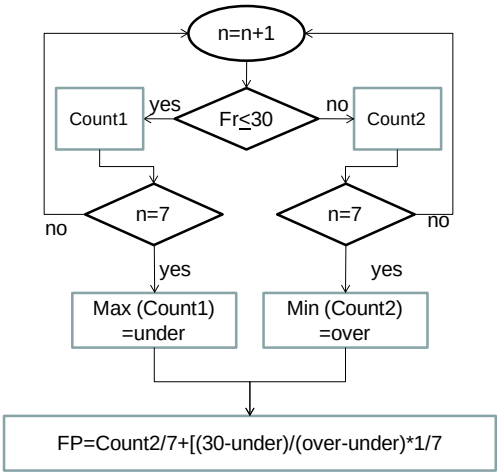


Figure 22. Flow diagram for calculating forecast probability for a hypothetical six member ensemble using the uniform ranks method. Count1 is the total number of ensemble members that meet the clear threshold (30%) and Count2 is the total number of members above the threshold.

Continuing with our example, a 6-member ensemble would have 7 verification bins. The observation would then have a probability of 1/7 in falling into any one of the 7 verification bins. Each ensemble member is evaluated against the 30% cloud-fraction threshold. Values less (greater) than the threshold represent each ensemble member's forecast probability of a clear (cloudy) event. The ensemble forecast probability for a clear event simply results from 1) summing the number of ensemble forecasts over and under the threshold, 2) dividing by n+1 bins (7) to calculate the cumulative ensemble forecast probability (.57), and 3) adding a factor that accounts for the probability of a clear event occurring between the highest member forecast below the threshold and the lowest member forecast above the threshold (.07). The resulting probability of clear conditions is 64% using the uniform ranks method.

The weighted ranks method which assigns probabilities based on past ensemble performance is generally expected to further improve ensemble probability forecasts. It is a simple calibration procedure. We previously showed that our ensemble lacks variability; thus, the rank histogram for this example was configured to match the U-shape of our results (Figure 23).

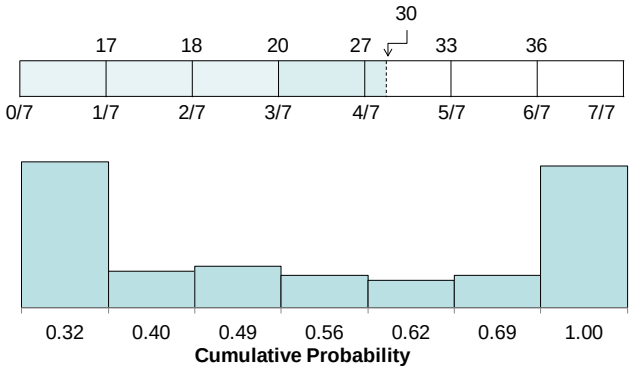


Figure 23. Weighted ranks method for calculating forecast probability. Probability calculation schematic using a hypothetical six member ensemble.

As in 1) and 2) of the uniform ranks method, we begin with the uniform cumulative probability 4/7. Next, we replaced this value with the cumulative probability of the rank histogram. The new probability .56 is essentially the probability of clear

conditions weighted by the expected performance of the model. Now, we calculate the added probability of clear conditions occurring between ensemble members 4 and 5.

$$\frac{30-27}{33-27}\times(.62-.56)=.5\times.06=.03 \qquad 25$$

Again, we assume the probability between members 4 and 5 is uniformly distributed. Finally, the probability for a clear event occurring based on previous model performance is 59%, the sum of the weighted probability (56%) and the added probability (3%).

The tails of the uniform and weighted ranks methods require special handling. We must account for occasions when all ensemble-member forecasts fall above or below the threshold. We use equation 26 for the former and equation 27 for the latter.

$$\frac{30}{m_1}\times p(bin_1) \qquad 26$$

or

$$p(bin_n)+\frac{(30-m_n)}{(100-p(bin_n))} \qquad 27$$

where m_n is the forecast value of the n^{th} ensemble member and $p(bin_n)$ is the cumulative probability of the n^{th} ensemble member. Here we also assume that the probability is evenly distributed amongst the calibrated bins.

These three methods of calculating ensemble probability will be compared to the mean and control forecasts. The democratic voting method assumes each ensemble member is equally likely to occur. The uniform ranks method produces a continuous probability, which accounts for the probability of an event occurring between ensemble members above and below the forecast threshold. The weighted ranks methods is a calibration applied to the uniform ranks method using historical forecast performance information.

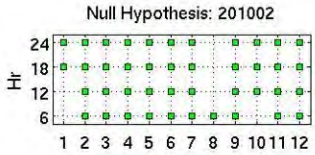
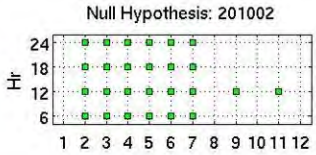
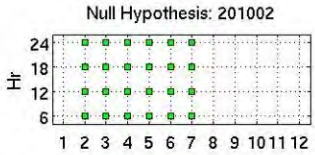
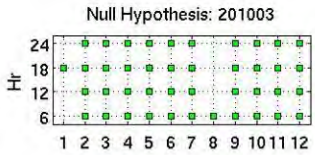
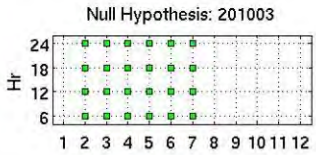
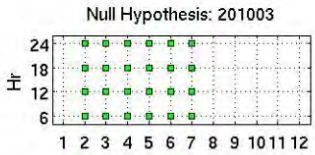
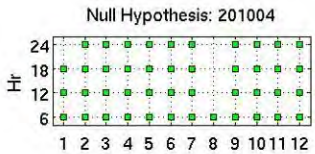
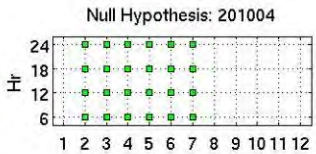
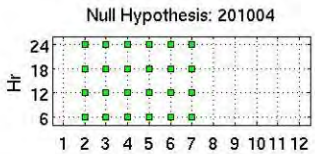
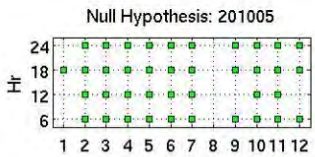
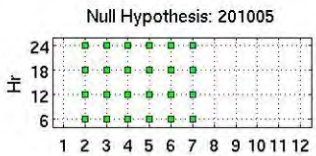
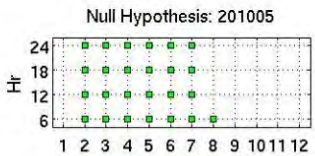
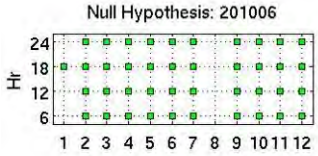
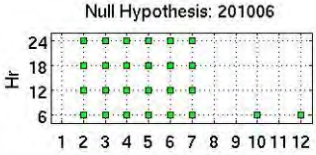
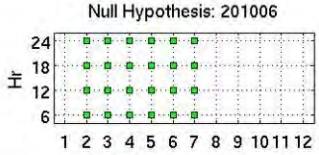
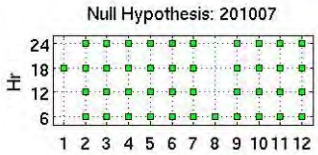
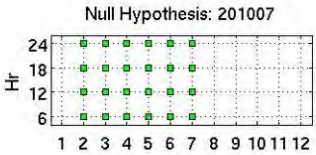
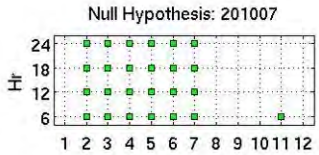
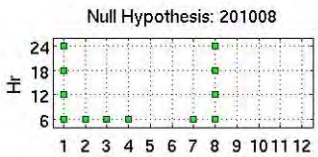
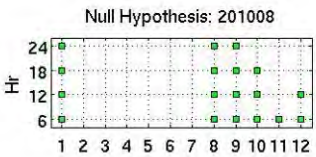
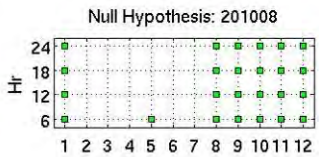
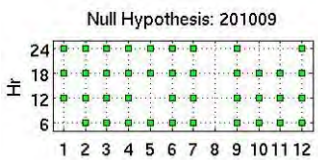
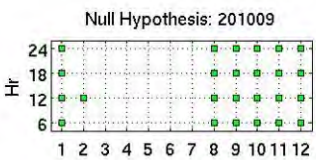
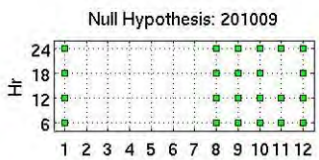
C. EVALUATING SEASONAL VARIABILITY

Understanding the seasonal variability of the model and analysis provides the context necessary to evaluate not only how the model performs in different cloud environments, but also how the model performs during atmospheric transitions and seasons. We begin seasonal characterization by examining the month-to-month variability in the probability distribution of the monthly verification rank histograms (VRH). Our distributions are not Gaussian, so the standard parametric tests are not particularly useful (Hogan et al. 2001). Instead we return to Kolmogorov-Smirnov's (K-S) two-sample test (Wilks 2006). We tested the null hypothesis that the VRH from one month will represent every other month within the data set (Figure 24). Each box indicates the acceptance of the null hypothesis that the performance of the ensemble in respective months on the x axis is significantly similar to the month identified as the "Null Hypothesis." The K-S tests are performed at forecast hours 6, 12, 18, and 24.

1. Ensemble

Results indicate that model performance in each region can be grouped into two groups, February–July/August and August/September–January. Applying K-S test in each region, we found that the model performs similarly for the months of February–July regardless of cloud cover frequency. A significant shift in the model performance takes place as the season shifts from summer to fall, but appears to stabilize by September in all three regions. Except for January and August, the cumulative distributions for Region 1 remained relatively consistent throughout the year. The cumulative distributions of Region 2 and Region 3, however, changed significantly after August. The results indicate that January best represents the months of August–December in both Region 2 and Region 3. This behavior coincides with modifications made to the ensemble in the July 2010 (Zhu 2010). We revisit this later, but for now we conclude that grouping months based on model performance or seasons is reasonable.

Kolmogorov-Smirnov Test 2

Region 1	Region 2	Region 3
		
		
		
		
		
		
		
		

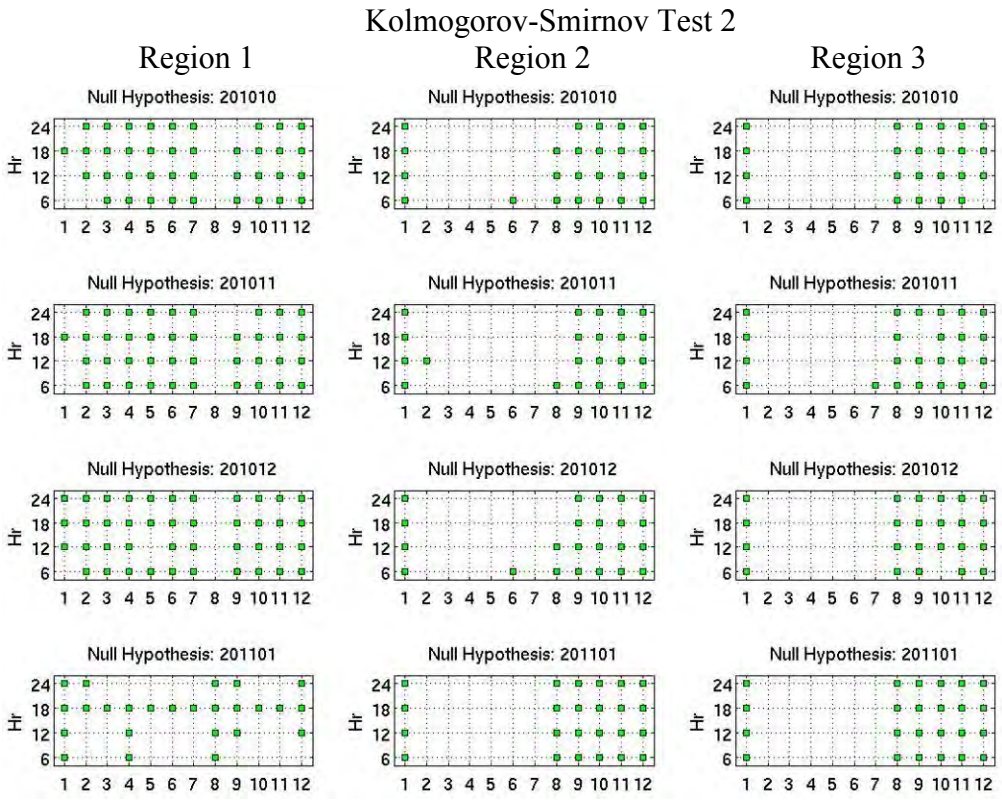


Figure 24. KS-test2 used to determine the seasonal variability in model performance. KS-test2 is performed for each forecast hour (left axis) and each month of the dataset (bottom axis). When the month/hour is marked with a square, the cumulative distribution of the verification rank histogram is similar to the distribution of the month identified as the Null Hypothesis.

2. Analysis

Our evaluations of the analysis mean indicated that the mean cloud-cover remains consistent throughout the data set, the two seasons identified in our K-S test results were divided into four seasons. Seasons 1 and 3 consists of two months, January/February and July/August, respectively, and are considered transitional periods. Seasons 2 and 4 consists of four months, March–June and September–December, respectively.

Since both GACE and WWMCA tend toward 0 and 100% cloud cover, we use these values to further analyze seasonal changes in WWMCA. We count the number of instances for which the fractional cloud coverage is 0% or 100% in each region (values between 0 and 100% cloudy are excluded). The results for the 0600 UTC analyses are shown in Figure 25, Figure 27, and Figure 28.

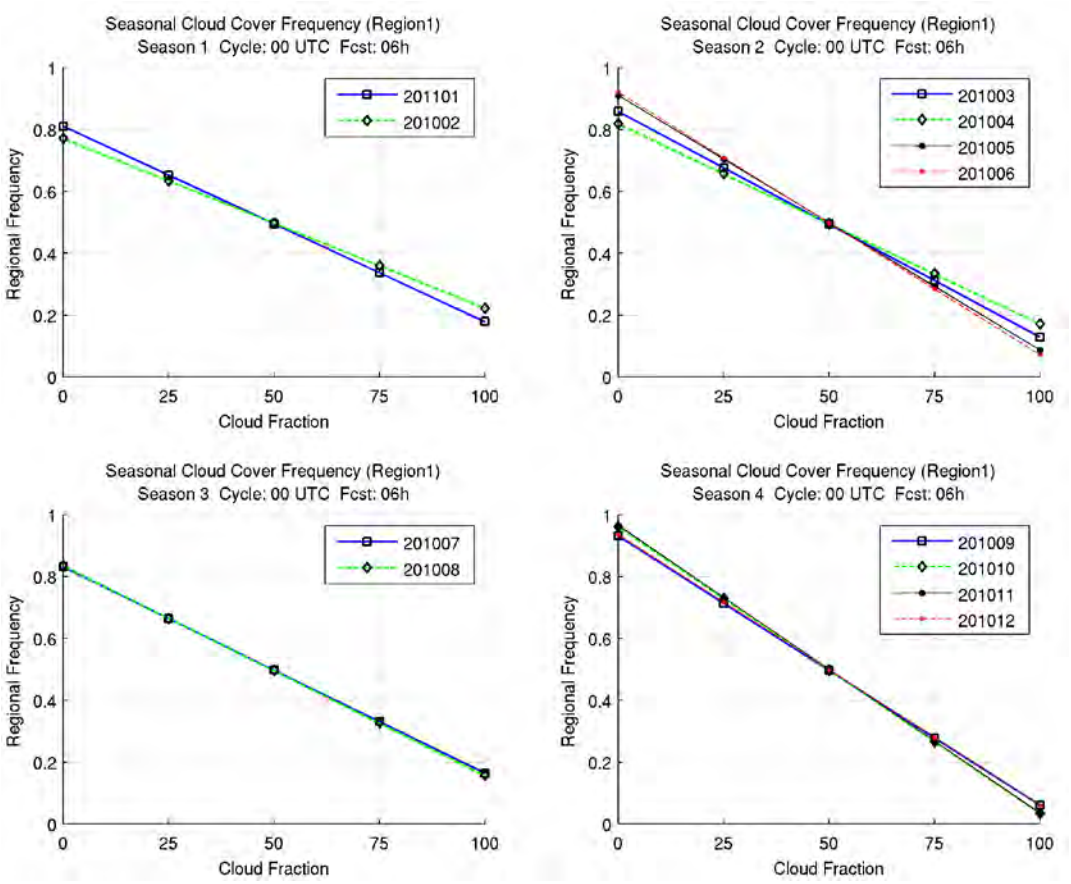


Figure 25. WWMCA cloudy vs. clear plot (Region1). Percentage of time 0% and 100% cloud fractions are observed. Data is divided into four seasons.

We gather from Figure 25 that the annual variability in cloud cover is minimal for Region 1. This is expected based on the criteria we used to select the region, but it is not the climatological norm. Normally, precipitation increases during the winter months making cloud-free observation less frequent. Figure 26, however, indicates that the mean outgoing longwave radiation (OLR) in Region 1 from September to December was higher than normal. The presence of clouds reduces the absorbed solar radiation and emitted longwave, or infrared, radiation of the surface and atmosphere (Barry and Chorley 1998). Therefore, anomalously high OLR values indicate abnormally low cloud cover.

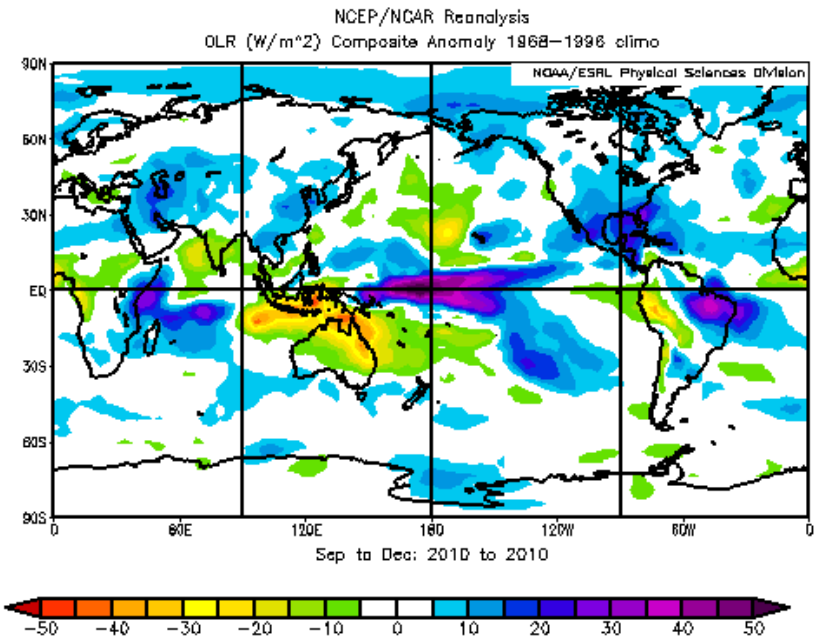


Figure 26. NCEP/NCAR Reanalysis. The impact of La Nina year is evident in the anomalously clear conditions over Region 1 and Region 2 during the months of September through December. (From Zhu 2011 or n.d.?)

The cloud cover distribution in Regions 2 shifts from season to season (Figure 27). Although cloudy conditions are prominent, cloud cover increases throughout season 2 with the onset of the monsoon (rainy) season. August marks the peak of the monsoon season and the transition to season 4. Cloud cover continues to decrease in season 4 until clear becomes the predominate condition. These results indicate that WWMCA cloud cover is consistent with well known, large-scale atmospheric conditions within the region.

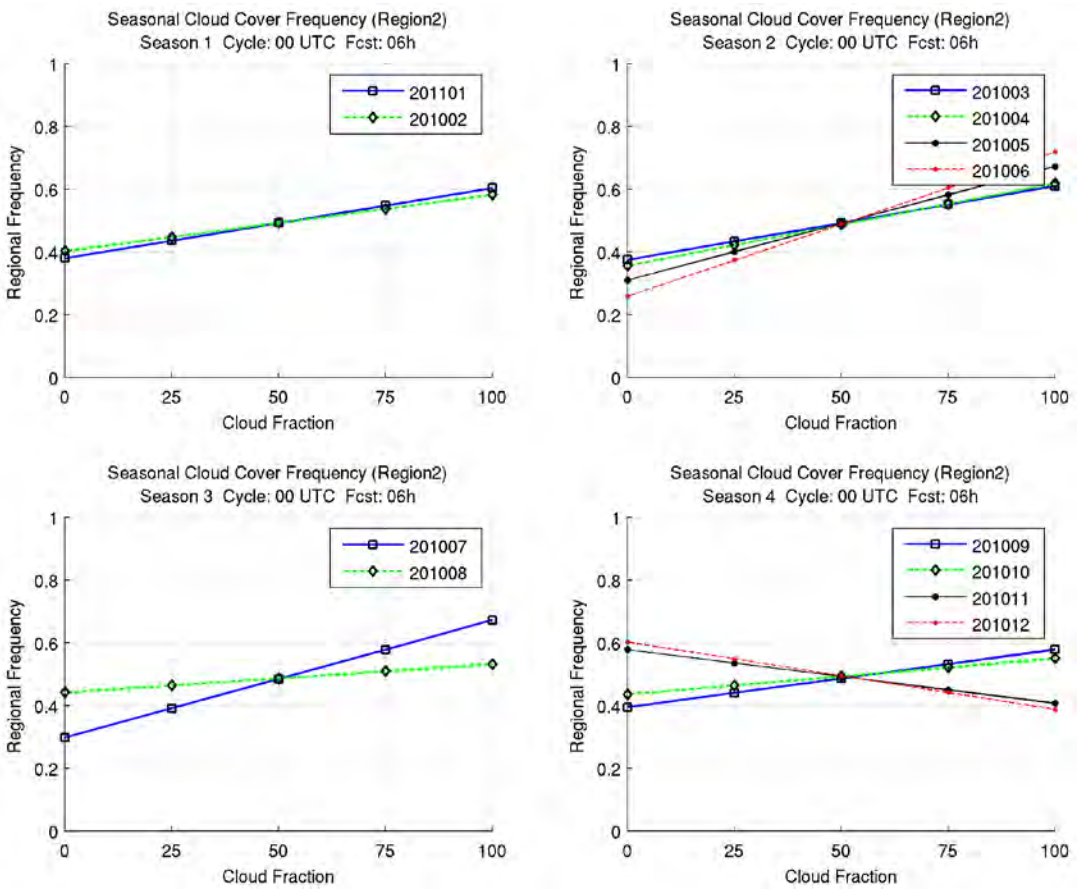


Figure 27. WWMCA Cloudy vs. Clear Plot (Region2). Percentage of time 0% and 100% cloud fractions are observed. Data is divided into four seasons.

Variability in Region 3 is driven by the oscillation of the Inter-tropical Convergence Zone (ITCZ). January, season 1, marks the most southern deflection of the ITCZ (with respect to Region 1) and is generally the cloudiest month. The decrease in cloudiness seen in season 2 indicates the northern deflection of the ITCZ. In August (season 2), the ITCZ reaches its most northern deflection and begins to return south (season 3) bringing increased cloud cover which peaks during season 4. Again, we can clearly see that WWMCA is fairly reliable in capturing large scale oscillations in cloud cover.

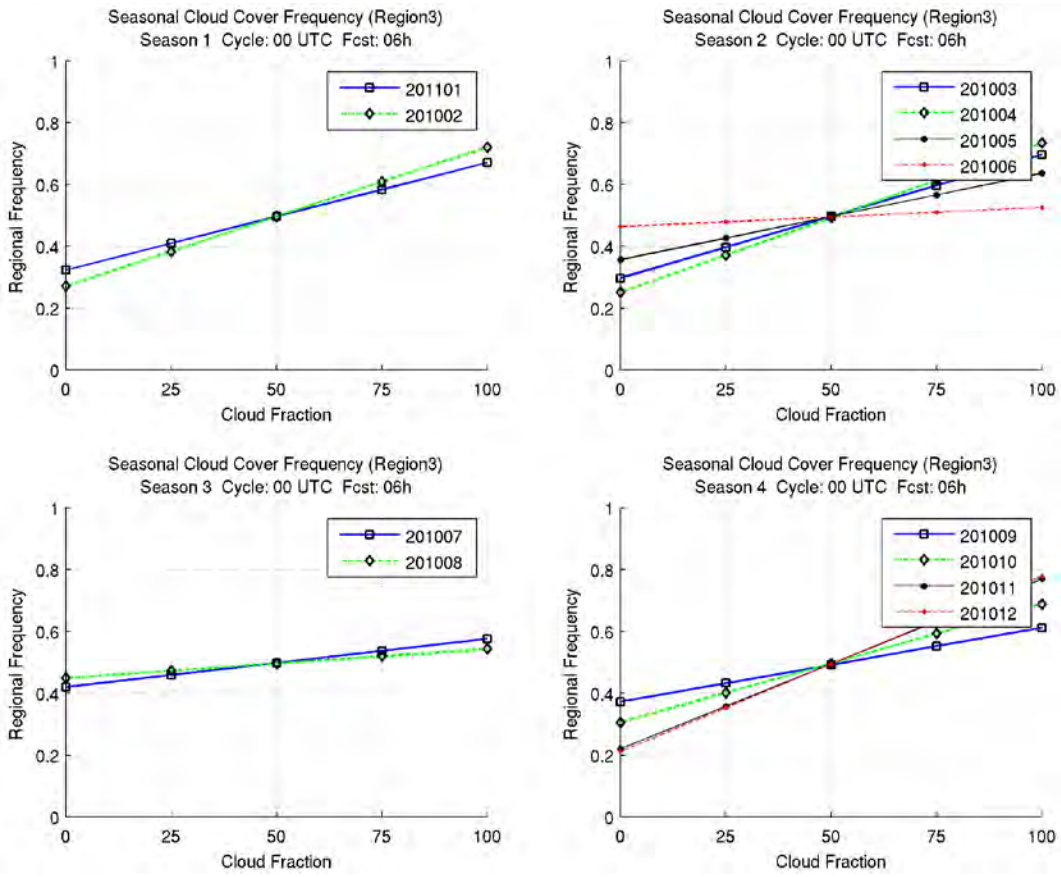


Figure 28. WWMCA Cloudy vs. Clear Plot (Region 3). Percentage of time 0% and 100% cloud fractions are observed. Data is divided into four seasons.

D. VERIFYING PROBABILITY FORECASTS

Forecast verification typically addresses how well the forecast limits the unwanted effects of "bad" weather. This defensive perspective is called an adverse weather perspective. Adverse weather centric forecast verification describes how well the forecast assists in the protection of assets, people and operations. Image collectors have the same goal but tend to judge forecasts based on operational success rather than forecast success. The primary focus of image collection forecasts is to assist in the collection of cloud-free imagery instead of protecting against the collection of cloud contaminated images. This redefinition is subtle but has significant implications for how one chooses to verify forecasts.

The operational viewpoint, or "preferred-weather" perspective, transforms and recasts the outcomes of the traditional 2x2 verification contingency table. As seen in Table 5, the nomenclature is preserved, but the definitions are adjusted to reflect the desire to collect cloud-free images. Collecting a clear image (CR) becomes the optimal result (HT) and collecting a cloudy image (MS) becomes the worst result (FA). Not collecting a clear image (FA) becomes a missed collection opportunity (MS), and not collecting a cloudy image (HT) is redefined as correctly rejecting a cloud filled image (CR).

Table 5. Transposition of an adverse weather perspective to a preferred weather perspective.

Outcome	Weather Perspective	
	Adverse	Preferred
Not Collect Cloudy Image	HT	CR
Not Collect Clear Image	FA	MS
Collect Cloudy Image	MS	FA
Collect Clear Image	CR	HT

		FORECAST/ACT		
		YES (clear)	NO (cloudy)	Total
OBSERVATION	YES (clear)	Hit (HT) (Clear Collection)	Miss (MS) (Missed Clear Opportunity)	Total Yes Obs (TYobs)
	NO (cloudy)	False Alarm (FA) (Cloudy Collection)	Correct Rejection (Reject Cloudy Collection)	Total No Obs (TNobs)

Figure 29. A 2x2 Contingency Table: Special case for cloud-free forecasts and observations using preferred-weather perspective.

Using the preferred-weather perspective, as illustrated in Figure 29, we verify and compare several forecast methods. Forecast verification includes ensemble probability forecasts (uniform ranks, weighted ranks, democratic voting) and ensemble-based deterministic forecasts (ensemble mean and control). Our analysis begins with commonly used verification measures, outlined in Table 6, for non-probabilistic forecasts.

Table 6. Deterministic forecast measurements derived from a 2x2 contingency table. Skill scores primarily calculated from the number of hits, misses, correct rejections, and false alarms. Expected (e) outcomes are used to modify the Heidke Skill Score. The hit rate and false alarm rate are used to calculate the True Skill Score.

Heidke Skill Score	
$HSS = \frac{HT + CR - eRF}{n - eRF}$ $n = HT + CR + MS + FA$	Correct Random Forecasts $eRF = eHT + eCR$
	HT Due to Chance $eHT = \frac{(HT + FA)(HT + MS)}{n}$
	CR due to Chance $eCR = \frac{(CR + FA)(CR + MS)}{n}$
True Skill Score	
$TSS = H - F$	Hit Rate: $H = \frac{HT}{HT + MS}$
	False Alarm Rate: $F = \frac{FA}{FA + CR}$

The hit rate (H), also known as the probability of detection, is the ratio between the number of correct clear forecasts and the total number of clear occurrences. The hit rate describes the ability of the forecast to predict when clear conditions will exist. The hit rate ranges from 0 to 1 with 1 being the perfect score. Using this metric alone causes problems because a forecast that prefers clear can have an excellent hit rate but an unacceptable number of false alarms.

The false alarm rate (F), also known as the probability of false detection, is the ratio between the number of incorrect clear forecasts and the number of cloudy occurrences. The false alarm rate provides an indication of how often the forecast incorrectly predicts clear conditions relative to the number of cloudy conditions observed. The false alarm rate should not be confused with the false alarm ratio, which is the fraction of clear forecasts that turn out to be incorrect. The false alarm rate ranges from 0 to 1 with 0 being the perfect score. This metric alone also causes problems because a

forecast that prefers cloudy conditions will have fewer false alarms, which reduces (improves) the score. In addition, the false alarm rate ignores misses which are a critical part of the forecast evaluation.

The utility of the H and F are often combined using Relative Operating Characteristic (ROC) diagrams (Figure 30). The ROC diagram provides a graphical representation of the forecast's ability to discriminate between dichotomous forecast decisions at different probability decision thresholds. It is a joint probability plot of $p(y_i,o_1)$ vs. $p(y_i,o_2)$, where y_i represents the probability decision threshold with i ranging from 0 to 1, and o_1 and o_2 are the hit rate and false alarm rate, respectively, at each probability decision threshold i.

When the probability threshold is 0 or at point (0,0) of the ROC diagram, clear conditions are always forecasted. As the probability decision threshold increases from lower to higher values, the number of yes forecasts decreases as does the hit and false alarm rates. At the highest probability decision threshold, (1,1) on the ROC diagram, the event is never forecasted. Each hit rate and false alarm rate pairing is connected from (0,0) to (1,1) to produce the forecast's ROC curve.

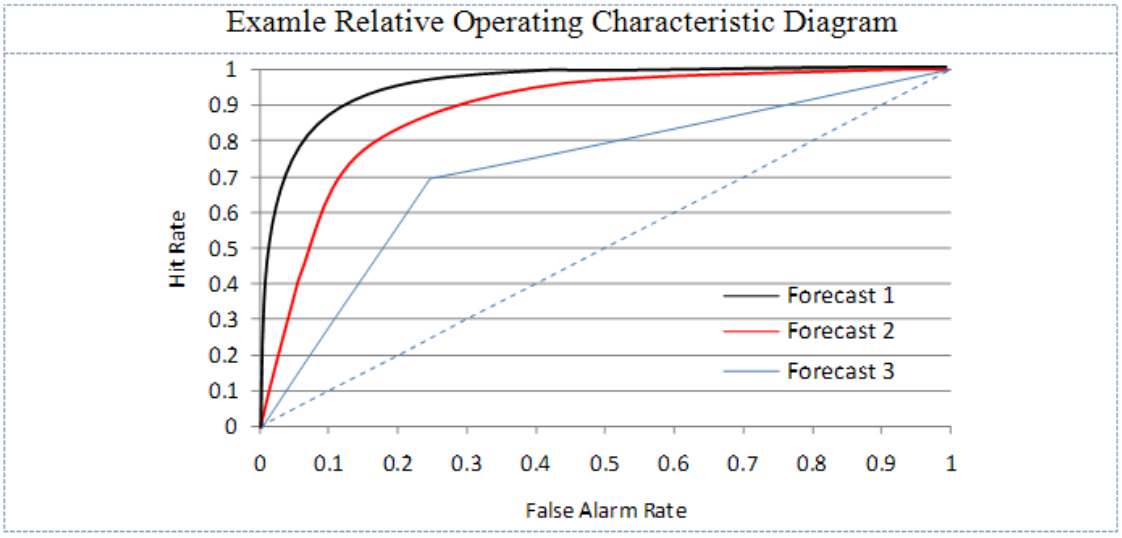


Figure 30. Example ROC Diagram. Forecast 1 (black) indicates superior skill over Forecast 2 (red) and Forecast 3 (blue). Forecast 3 exemplifies a deterministic ROC curve that has little to no variation between member forecasts.

The area under ROC curves indicates the skill ensemble in distinguishing between clear and cloudy events. These curves also provide a quick and easy way to compare forecasts. Forecasts with ROC curves closest to point (0,1) demonstrate the most skill. Forecast that produce ROC curves that are entirely above and to the left of other forecast curves imply statistically significant dominance to all rational forecast users (Wilks 2006).

The curvature displayed in the ROC curves is an indication of the variability in skill demonstrated by the forecast across probability decision thresholds. Therefore, forecast verifications that produce the same 2x2 contingency table despite differing user sensitivities do not produce typical ROC curves. All hit and false alarm rates converge at a single point on the diagram. This is an undesirable signature for ensembles. It is an indicator that the ensemble behaves deterministically and that there is insufficient variability between member forecasts.

The Heidke Skill Score (HSS) is a more complete verification measure. It accounts for hits, misses, false alarms, and correct rejections. The HSS measures the success of the forecasts after removing the success resulting from random chance. The marginal probability of clear (cloudy) forecasts and observations are multiplied to obtain the probability of a correct clear forecast by chance,

$$pHT = \frac{(HT + FA)}{n}\frac{(HT + MS)}{n} = \frac{(HT + FA)(HT + MS)}{n^2} \tag{28}$$

$$pCR = \frac{(CR + FA)}{n}\frac{(CR + MS)}{n} = \frac{(CR + FA)(CR + MS)}{n^2} \tag{29}$$

$$eRF = \frac{(HT + FA)(HT + MS)}{n} + \frac{(CR + FA)(CR + MS)}{n} = eHT + eCR \tag{30}$$

where *pHT* is the probability of a correct clear forecast by chance, *pCR* is the probability of a correct cloudy forecast by chance, and *n* is the total number of forecasts. The parameters *pHT* and *pCR* are multiplied by the total number of events to produce the expected number of hits (*eHT*) and expected number of correct rejections (*eCR*),

respectively. The sum of *eHT* and *eCR* defines the expected number of correct random forecasts (*eRF*). This value is then subtracted from the correct number of forecasts to calculate the Heidke Skill Score.

$$HSS = \frac{HT + CR - eRF}{n - eRF} \qquad 31$$

In meteorology, persistence or climatology is generally used to represent *eRF*. The HSS, as presented, penalizes forecasts for performing well. Table 7 shows that defining chance based on the outcomes of a forecast's 2x2 contingency table masks true skill differences in one-to-one forecast comparisons. The two forecasts only differ in the number of hits and misses. Forecast 1 (**bold**) has more hits and Forecast 2 (*italic*) has more false alarms. Forecast 1 performs better, but the improved performance is attributed to chance. Therefore, Forecast 1 has a larger number of forecasts attributed by chance than Forecast 2. This problem is not critical because HSS is normalized by a factor of $n - eRF$ and Forecast 1 receives the highest skill score of the two forecast methods.

Table 7. Forecast outcome comparisons. Forecast 1 produces more hits by chance than Forecast 2, and Forecast 2 produces more correct rejections by chance although forecasts are made under the same conditions. This highlights the need to use climatology instead of contingency table calculations of eRF.

		Forecast 1		Forecast 2			
		clear	cloudy	clear	cloudy	eHT	
OBSERVATION	YES (clear)	**200**	**20**	*190*	*20*	**eHT=193.6**	*eHT=184.8*
	NO (cloudy)	**20**	**10**	*30*	*10*	**eCR=3.6**	*eCR*=4.8

We choose to use the frequency of cloud cover, however, to ensure that all forecasts are evaluated equally. The number of correct forecasts expected by chance is redefined by substituting *eRF* with equation 32 that accounts for the number of hits and correct rejections based on the climatological probability of clear and cloudy events.

$$eRF = HT * p(clear) + CR * p(cloudy) \quad (32)$$

The Heidke Skill Score ranges from -∞ to 1 with 1 being the perfect score. Values less than or equal to 0 indicate that the forecast is worse than random. A weakness of this score is that it tends toward zero for rare events (Hogan 2009). This makes it unreliable in regions where clear collections are most desired. In addition, choosing to use climatology to represent *eRF* reduces the impact of misses and correct rejections on forecast skill.

The true skill statistic (TSS) (Wilks 2006; Flueck 1987)—also referred to as the Hanssen-Kuipers skill score—is the difference between H and F.

$$TSS = H - F \quad (33)$$

The score is useful in that it emphasizes the undesired outcomes of the 2x2 contingency table. The TSS is positive as long as H>F and equals 1 (perfect) when MS=FA=0. The true skill statistic accounts for forecast errors due to omission and commission. Always forecasting clear or cloudy yields a score of 0 (no skill) because MS=CR=0. Forecasts that are worse than random forecasts receive negative TSS values. Correct rejections (hits) improves the score more when clear conditions are persistent (rare).

The Odds Ratio Skill Score is another means of evaluating forecast skill. It has its basis in the odds ratio (OR), the probability that an event will occur vs. the probability that the event will not occur.

$$OR = \frac{\left(\frac{H}{1-H}\right)}{\left(\frac{F}{1-F}\right)} = \frac{HT * CR}{MS * FA} \quad (34)$$

The odds ratio ranges from 0 to ∞. When the score is equal to 1, the events being compared are equally likely to occur. Scores less than 1 (F ≥ H) indicates that the forecast has no skill; and because it is unbounded, the perfect score is ∞. The Odds Ratio is not as sensitive to hedging as with hit and false alarm rates. In addition, it provides a better assessment for rare events (Stephenson 2000). This score should only be used

when all outcomes occur at least once. If MSs or CRs equal zero, H and F equal one, respectively, and the odds ratio is undefined. The odds ratio is also undefined when FAs equal zero.

The odds ratio in equation 30 can be rewritten as,

$$OR = \frac{HT/FA}{MS/CR} \tag{35}$$

Writing the odds ratio this way helps us see that it is essentially the odds of collecting a clear image vs. the odds of missing a clear image. The Large positive numbers are preferred. Positive numbers mean that we are more likely to collect a clear image than to miss one. Positive values are expected in a persistently clear region, but its efficacy in persistently cloudy regions makes it an attractive measure of skill for this research.

In this research, we use the odds ratio as a skill score,

$$ORSS = \frac{OR - 1}{OR + 1} = \frac{HT * CR - MS * FA}{HT * CR + MS * FA} \tag{36}$$

The Odds Ratio Skill Score (ORSS) is mainly used in medical statistics, but has valid application in meteorological forecast verification (Stephenson 2000). A score of 0 is an indication of a random forecast. A perfect score is 1, but the forecast does not have to be perfect to attain it. If either misses or false alarms are zero, then the forecast gets a perfect score of 1. We avoid this problem by summing the respective outcomes within each region to produce a monthly 2x2 contingency. This fully populates the contingency table and increases confidence in the results of our ORSS calculations.

E. UTILITY OF OUTCOMES

We rely on the three assumptions discussed in the background section to assign reasonable inherent utility value to the outcomes. We note that clear preferences exist among the four possible outcomes (Assumption 1) and the preferences are transitive (Assumption 2). A rational operator will prefer HTs to CRs, CRs to MSs, and MSs to FAs. Correctly forecasting clear conditions will result in highest operator satisfaction. Correctly forecasting cloudy conditions will also result in high operator satisfaction, but

does not result in mission success so its utility is somewhat less than one – this result is modeled since CRs are only confirmed when the operator chooses not to follow the forecast. Operators will be dissatisfied with MSs and FAs because both constitute mission failure.

Assigning a utility value to HTs, FAs, and MSs is straight forward. The best outcome of cloud-free image collection operations is collecting a clear image (HT). The worst outcome is collecting a cloudy image (FA). Hits result in maximum operator satisfaction and FAs result in a cloudy image and the least operator satisfaction. Thus, we assign a value of 1 (maximum utility) to HTs and a value of 0 (minimum utility) to FAs. Because not collecting an available clear image is also undesired, MSs are assigned a utility value of 0. The inherent utility of HTs, FAs, and MSs define the upper and lower bounds of our utility function.

This leaves the utility of correct rejections as the lone utility value to be assigned by respective users. Risk-tolerant users are not impacted significantly by the collection of cloud filled imagery and are willing to collect even when the possibility of obtaining a cloud-free image is low. These users might assign a utility value of 0.2–0.3 to correct rejections. Risk-averse users are highly sensitive to the collection of cloudy imagery. Imagery processing may be costly; the desired imagery may be time sensitive; etc. These users might assign a utility value of 0.6–0.9 depending on the impact of cloudy collections on cloud-free operations. Users who find zero utility in correctly rejected cloud filled imagery will essentially collect regardless of the cloud cover and render the forecast impotent in collection operations.

Figure 31 shows the variability in the value a user can expect to obtain using perfect information. Each line represents a user defined correct rejection utility. Users who choose a utility value of 0 in correctly rejection cloudy imagery determine a cloud-free forecast to be of no consequence to their operation. Although situations may exist where the desire to collect clear imagery equals the desire to correctly reject cloudy imagery, we do not consider this case. Therefore, utility values range from 0.1 to 0.9. Based on the parameters of our cloud-free problem, expected utility value is bounded by 0 and 0.45. Expected value of perfect information peaks at p(clear)=50% where

uncertainty is highest. This peak coincides with the expected value of low tolerant users who assign 0.9 to the utility of correct rejections. Even high tolerant users, who find little value in avoiding cloudy imagery, have the potential to decrease their operational cost by 10%, and 5% at the most uncertain locations.

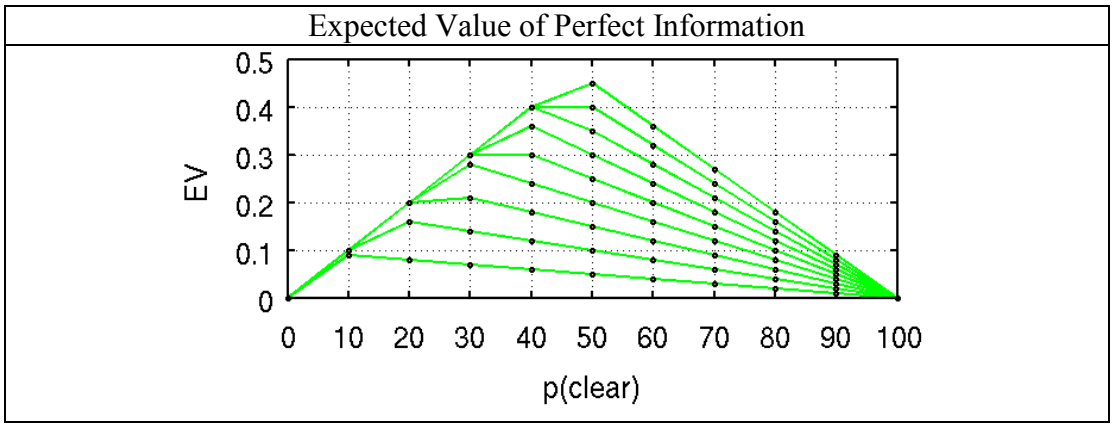


Figure 31. Example expected value chart (Perfect information). Each line represents the maximum expected value gain with the introduction of perfect information based on user sensitivity to correct rejections and probability of clear conditions. Utility values 0.1–0.9 are represented. Expected value increases as cloud cover frequency decreases and utility of correct rejections increase.

Figure 32 is an example of the expected value of partial or imperfect information. The red dotted lines represent the most sensitive user and the black dotted lines represent least sensitive users. The lines marked with yellow squares and green lines represent users between these extremes. Unlike perfect information, the value gained through the use of imperfect forecast information depends on the ability of the forecast to distinguish between events and non-events, also known as the percentage of correct forecasts (PC).

Notice that a forecast that is never correct (top left) has the same curve as perfect information (bottom right). This arises from the symmetric manner in which the problem is framed. Utility is maximized in uncertain environments and minimized when confidence in event occurrence or non-occurrence is high. Furthermore, the forecast that is always correct can potentially provide as much utility as a forecast that is always incorrect. A forecast that is certainly *incorrect* provides perfect information. The user of this information has but to do the exact opposite of what the forecast guidance suggests.

This mirroring attribute of expected value is symmetric about 50% probability of clear. The expected value of the forecast decreases as uncertainty decreases. The decrease, however, is not symmetric. As forecast uncertainty shifts from 1 to .5, mid-level users obtain greater potential value from forecasts than extremely sensitive users when the probability of clear is low. We also note that clear forecasts are not expected to add value when the probability of success is at or below the climatological probability of success.

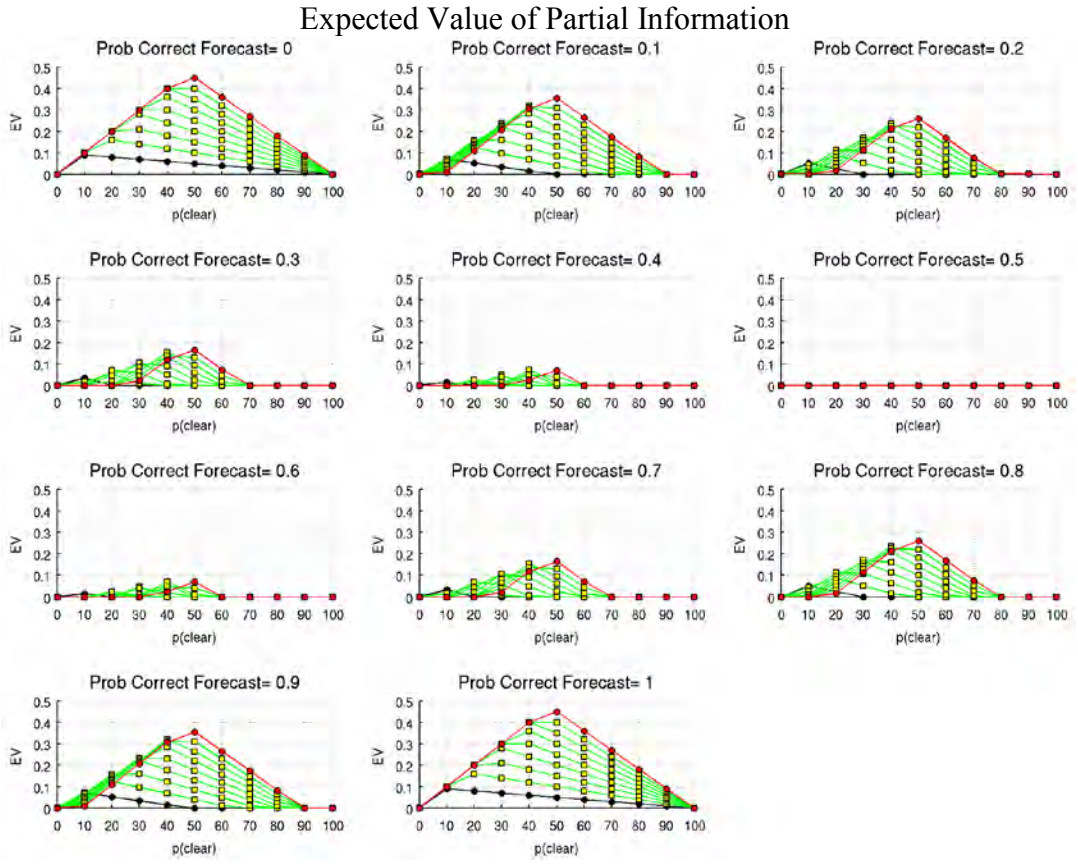


Figure 32. Example expected value chart (Partial information). Each line represents the maximum expected value gain with the introduction of imperfect information based on user sensitivity to correct rejections and probability of clear conditions. Utility value .9 (red dotted lines), utility value .1 (black dotted line), and utility values between .1 and .9 (yellow boxed lines) are plotted.

V. ENSEMBLE SKILL

In this section, we cover a few of the most interesting cloud-free forecasting results that arise using the global advection cloud ensemble. Our analysis includes several different data groupings. Herein, we evaluate the skill of cloud-free forecasts over a 12-month period. The skill of a forecast could reasonably vary with a user's decision threshold, cycle or forecast hour, and region. In addition, forecast skill can fluctuate with the dynamics and/or seasons of a region.

The skill evaluations are centered on five basic ensemble forecast techniques. We examine the skill of the ensemble mean and control as deterministic forecast methods. Although the ensemble mean forecast carries with it information about uncertainty in cloud cover amount, the uncertainty information cannot be directly communicated. Probability forecasts communicate forecast uncertainty in a way that allows users to manage their risk based on a probability decision threshold (DT). For this application of ensemble information, we use the democratic voting, uniform ranks, and weighted ranks for our skill evaluations. Each forecast method is verified independently, and skill is assessed.

The ensemble forecasts are verified globally at each grid point. Each forecast method is compared to WWMCA to generate a 2x2 contingency at each grid point. Deterministic forecasts are not sensitive to user decision thresholds, so the ensemble control and mean forecasts are verified against WWMCA using the 30% cloud fraction threshold independent of probability decision thresholds. Ensemble probability forecasts are verified using probability decision thresholds from 10%–90% at 10% intervals and WWMCA (Table 8). Verification outcomes at each grid point are summed monthly for each forecast method. The resulting monthly contingency tables are then used to calculate the ensemble skill for each forecast method.

Table 8. Probability forecast verification. Hits, false alarms, misses and correct rejections are defined relative to forecast probability (*p(clear)*), probability decision threshold (DT), analysis value (WWMCA), and cloud fraction threshold (30%).

HT	$p(clear) >= DT$	&	$WWMCA <= 30\%$
FA	$p(clear) >= DT$	&	$WWMCA > 30\%$
MS	$p(clear) < DT$	&	$WWMCA <= 30\%$
CR	$p(clear) < DT$	&	$WWMCA > 30\%$

Our election to use climatological probabilities, which vary by grid point, to define chance rather than contingency tables makes it undesirable to calculate regional mean skill score. To do so, outcomes of the contingency table would be totaled relative to probability decision thresholds (Appendix B) and a regional mean cloud cover is required to construct the regional HSS. This method is undesirable because it does not account for variations in skill with differing cloud cover frequencies within the regions. This is particularly the case in Region 1 where cloud cover frequencies at northern grid points can be significantly different than southern points. The regional mean skill score, at least in the case of the HSS, can mask important sub-regional skill information, so we calculate the mean skill of each grid point within the region instead.

Monthly HSS and TSS values are calculated from the monthly contingency tables. At grid points where neither hits nor correct rejections are recorded, HSS is undefined. We set the HSS to zero to avoid computational errors in these cases. Likewise, when forecasts fail to record hits or false alarms, the hit and false alarm rates are undefined respectively. Thus, we set the hit or false alarm rate to zero and calculate the TSS. The mean skill score of each ensemble forecast method is then calculated with respect to user sensitivity to cloud cover, forecast hour, and cloud cover frequency. In our discussions of these skill score groupings, we often return to the contingency table outcomes, which have been summed relative to probability decision thresholds and the frequency of cloud cover within each region (Appendix B and Appendix C).

The skill of each ensemble forecast method is plotted on a series of charts. With these charts, we attempt to group our forecast results in ways that are compact,

meaningful and easily consumable. On these charts, the ensemble control forecast is represented with a plus (but included in the figure legends). The ensemble mean forecast (EMean) is represented by a blue square. The democratic voting method (DVote) is represented by a green diamond. The uniform ranks method (URank) is represented by a black dot. The weighted ranks method (WRank) is represented by a red dot. We evaluate the skill of these methods simultaneously.

1. DECISION THRESHOLD VARIATIONS

We begin our regional evaluations of the ensemble by examining forecast skill relative to users with different sensitivities to cloud cover. This measure is synonymous with a user's priority for a clear collection. Lower (Higher) decision thresholds suggest that the image is of the highest (lowest) priority; users are willing (reluctant) to accept significant risk to collect the image. For each method and region, the spatial mean and standard deviation of the HSS and TSS are calculated and binned by probability decision thresholds. The resulting values are plotted monthly. In this section, we show the results from the 0000 UTC cycle and 6-h forecast.

1. Region 1 (Persistently Clear)

Figure 33 shows the HSS results for Region 1. The overall skill of the ensemble does not change significantly between forecast methods. Differences and shifts in skill correspond to seasonal changes that are commonly observed within the region. Differences are more prominent in some months than others, but the overlapping of the error bars (standard deviation) suggests that no significant difference exist. We focus on the differences in the mean skill with the understanding that the conclusions may not apply in all cases.

From February to April the skill decreases for each ensemble forecast method. Probability forecasts show the most distinction in skill between probability decision thresholds. In February, all forecast methods remain in fairly good agreement below 50% probability of clear conditions, but the ensemble probability forecasts are markedly better. By April, however, the differences are again minimal. The major reason for the improvement in skill of the probability forecasts over the ensemble mean and control

forecasts below 50% probability of clear arises primarily from the ratio of hits to correct rejections. The ensemble mean forecast has more correct rejections but fewer hits and more misses at these grid points. Because all forecasts produce a significant number of correct rejections, the HSS of the ensemble mean and control, which produce more misses, are reduced as compared to the probability forecasts.

Examining the TSS (Figure 34), however, helps us see that the high number of hits come at the cost of more false alarms for the probability forecasts. Therefore, the relative skill of the deterministic and probability forecasts is reversed below 50% probability of clear. We also note that in April there's a separation in skill among the ensemble probability forecasts. Here the weighted ranks method forecasts clear more often and produces more false alarms than the democratic voting and uniform ranks methods.

Probability forecasts demonstrate inferior skill as compared to deterministic forecasts above 60% probability of clear conditions during the months of February through April. The worst HSS at these decision thresholds is achieved by the weighted ranks method. At these probabilities, the weighted ranks method always forecasts cloudy, so only misses and correct rejections are recorded. Although the other probability forecasts perform better, the ensemble mean and control demonstrate the most skill at higher probabilities.

The TSS confirms the inferior skill of the weighted ranks method with extremely low tolerant users (90% DT). Whereas the HSS attributes skill to the weighted ranks method for correct rejections, TSS attributes zero skill for forecasts that always forecast clear or cloudy conditions. All other ensemble forecasts produce similar skill at high probability decision thresholds for February, March, and April.

Looking at the next few months, the ensemble forecasts perform similar to the previous three months. During the months of May and June, the HSS of probability forecast is slightly better than deterministic forecasts below 50% probability of clear and worse above 60% probability of clear and reverses with the TSS. The skill score values

are also comparable to those seen in the previous months. The relatively high skill during these months is primarily due to transient clouds that accompany frontal system as they pass through the region.

There is a decrease in HSS and TSS during the months of July through September as the southeastern quadrant experiences increased cloud cover brought on with the South Asian Monsoon. The HSS suggests that the ability of the forecasts to distinguish between cloud and clear is diminished significantly. The decrease in the TSS is not as dramatic, but the error bars indicate that skill varies significantly within the region. However, the HSS and TSS of all ensemble forecast methods are comparable. In addition, we no longer see the skill disparity, at higher decision thresholds between the weighted ranks method and other ensemble forecast methods.

The lack of differences in skill demonstrated among ensemble forecast methods persists from October through January. During these months, all ensemble forecasts behave deterministically, or demonstrate similar skill at each decision threshold. The weighted ranks method is the one exception. The weighted ranks method again demonstrates inferior skill at the 90% probability decision threshold. This result is upheld by both the HSS and TSS.

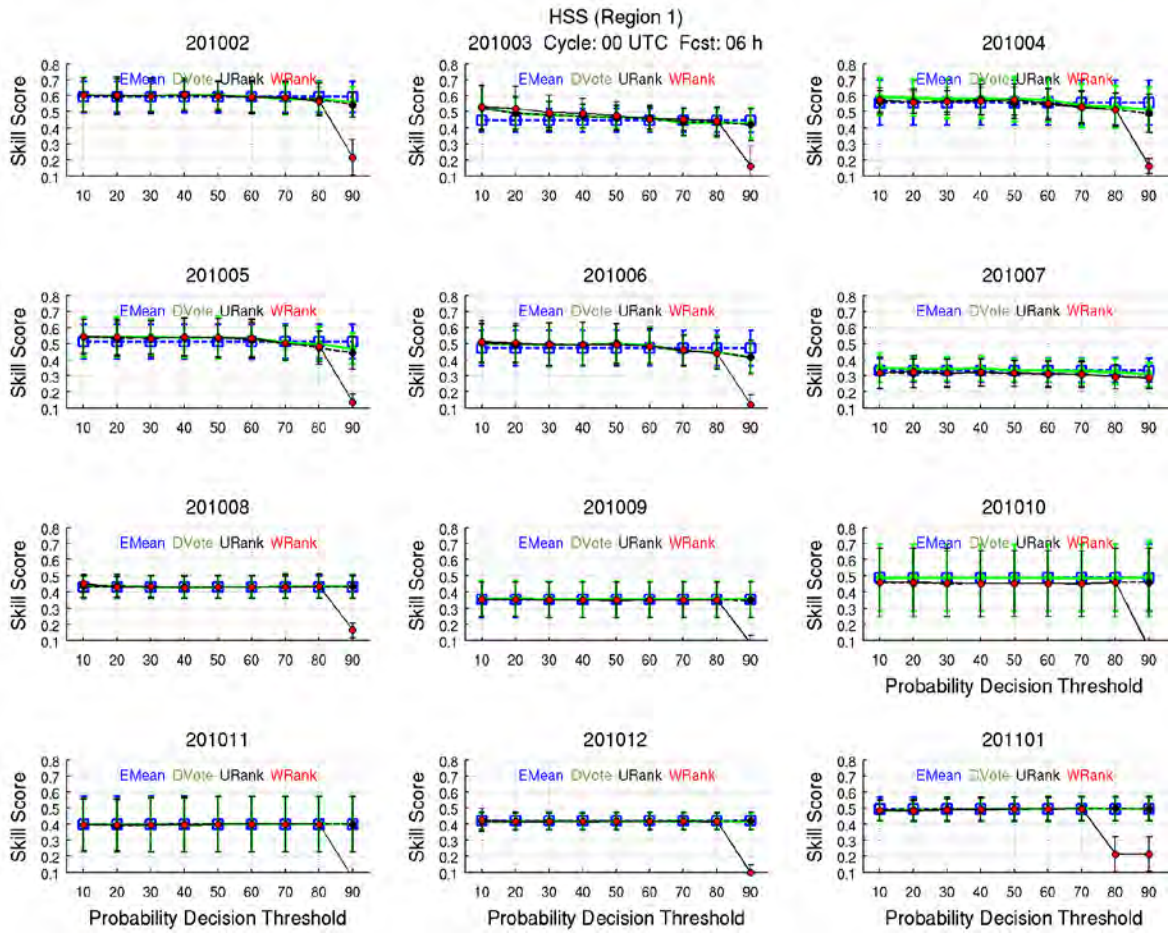


Figure 33. Heidke Skill Score relative to probability decision thresholds (Region 1). Ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) Heidke Skill Scores (left axis) compared to the probability decision threshold (bottom axis). Standard deviations in skill are represented with error bars.

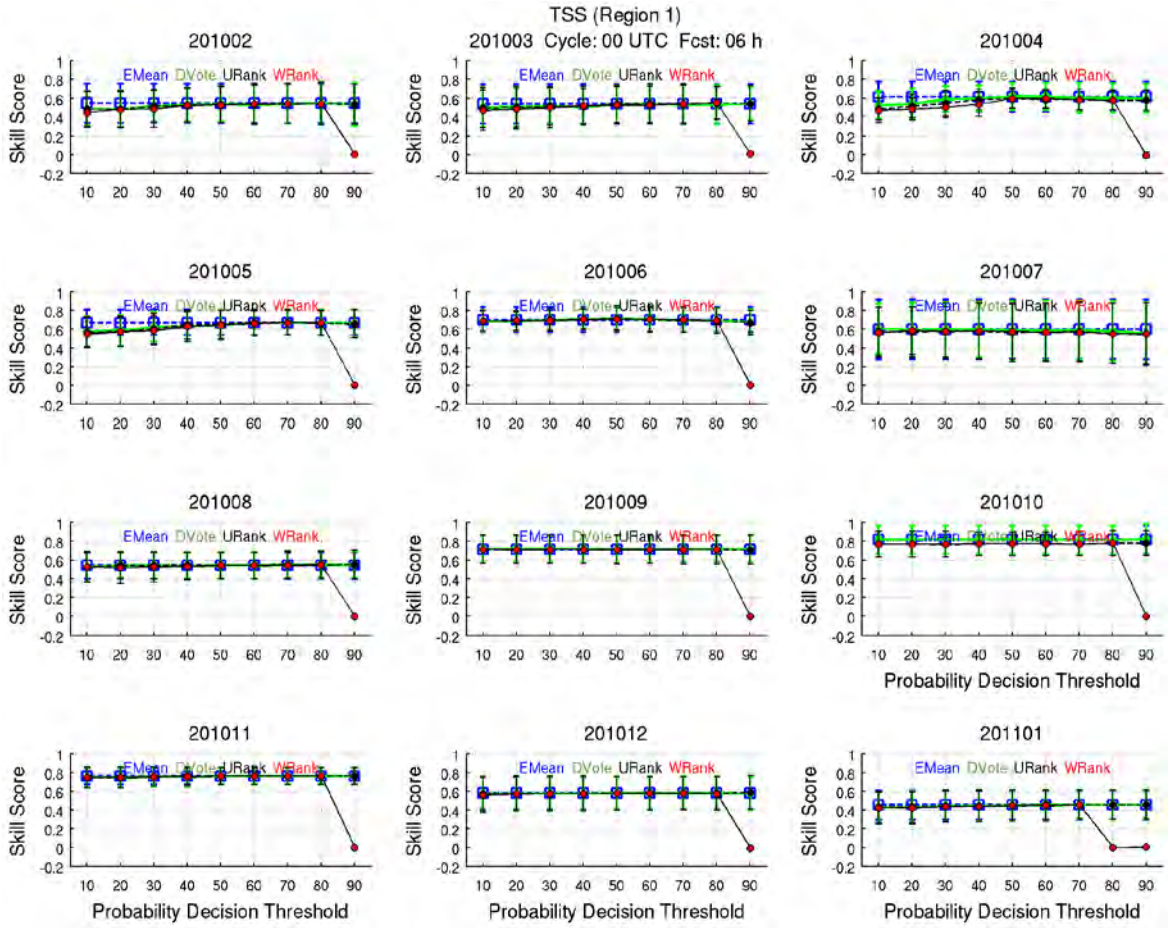


Figure 34. True Skill Score relative to probability decision thresholds (Region 1). Ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) True Skill Score (left axis) compared to the probability decision threshold (bottom axis). Standard deviations in skill are represented with error bars.

2. Region 2 (Variable Cloud Cover)

Figure 35 and Figure 36 serve as our focal points as we examine the skill of the ensemble in Region 2. As in Region 1, we see that skill differences in the ensemble forecast methods only occur with high and low tolerant users. More specifically, differences primarily occur below 40% and above 80% probability of clear. The error bars here also suggest that little difference exist between forecast methods even at the extreme decision thresholds. Never-the-less, comparing the mean skill of the forecast methods can help highlight forecast tendencies, weaknesses and strengths.

In February and March, there are distinct differences between the HSS and TSS of deterministic and probability forecast methods. Below 40% probability of clear, the weighted ranks method produces the lowest HSS and TSS of all other forecast methods. The primary factor in the reduction of skill is the tendency for the weighted ranks method to forecast clear at lower decision thresholds. This tendency produces fewer correct rejections than the other forecast methods. Cloud cover uncertainty within the region also causes the ensemble to consistently produce members that incorrectly forecast clear conditions as evidenced by the reduction in skill of democratic voting and uniform ranks methods as compared to the ensemble mean and control.

The next notable differences in skill, between the deterministic and ensemble forecast methods, occur above 80% probability of clear. The weighted ranks method persistently demonstrates inferior skill above 80% probability of clear. The skill, as demonstrated by both HSS and TSS, is significantly degraded because of the lack of clear forecasts at high probability decision thresholds. Deterioration in skill, as compared to deterministic methods, is also seen in the democratic voting and uniform ranks methods but not nearly as prominent as with the weighted ranks method.

In April, probability forecasts exhibit to show improvement in skill over the previous months. The distinct indication of a clear tendency below 40% and cloudy tendency above 80% can still be seen in the HSS. Here, however, probability forecasts are not just as skillful as the ensemble mean but more skillful between 30%–60%. The TSS shows that the uniform ranks and weighted ranks methods are far more skillful than democratic voting method as well. This improvement of skill with the TSS is extended to

all users except the most intolerant users (>80%). However, the increased size of our error bars leads us to believe that the improved skill of the ensemble may a function of a smaller forecast sample size in April.

From May through July, the HSS and TSS of the ensemble forecasts are similar to those seen in February and March. The probability forecasts preference clear conditions below 40% probability of clear and cloudy above 80% probability. Otherwise, all ensemble forecasts demonstrate equal skill.

Although Region 2 is characterized by significant variability in cloud cover, the ensemble forecast performance remains relative consistent throughout the year. Users who rely on the forecast in this region can expect strong similarities in ensemble skill between February–March and May–July. Forecast skill in April appears to be anomalous, but we have insufficient data to examine this further. Therefore, we accept the results with caution. The ensemble behaves deterministic throughout the rest of the forecast period, but skill remains evident.

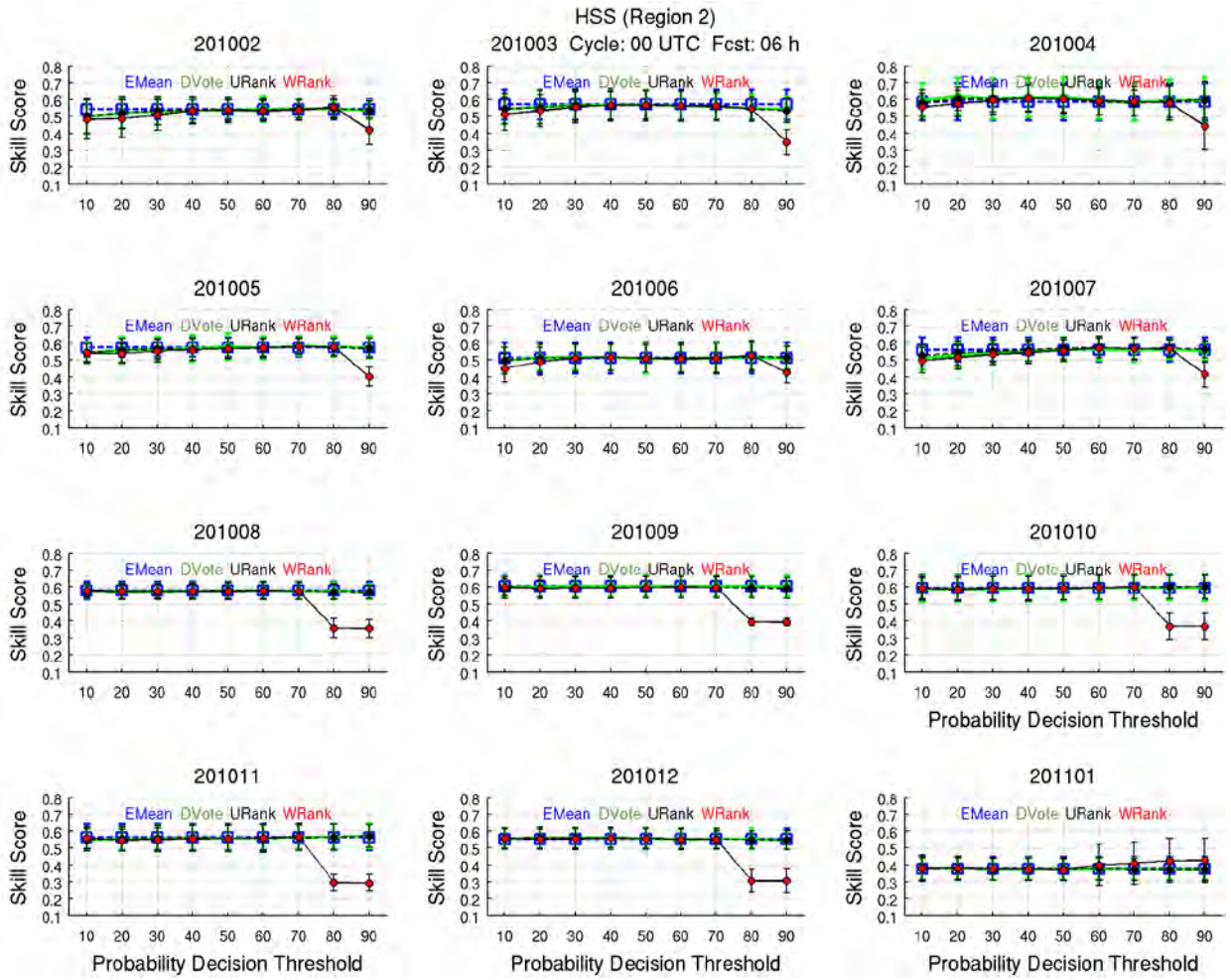


Figure 35. Heidke Skill Score relative to probability decision thresholds (Region 2). Ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) Heidke Skill Scores (left axis) compared to the probability decision threshold (bottom axis). Standard deviations in skill are represented with error bars.

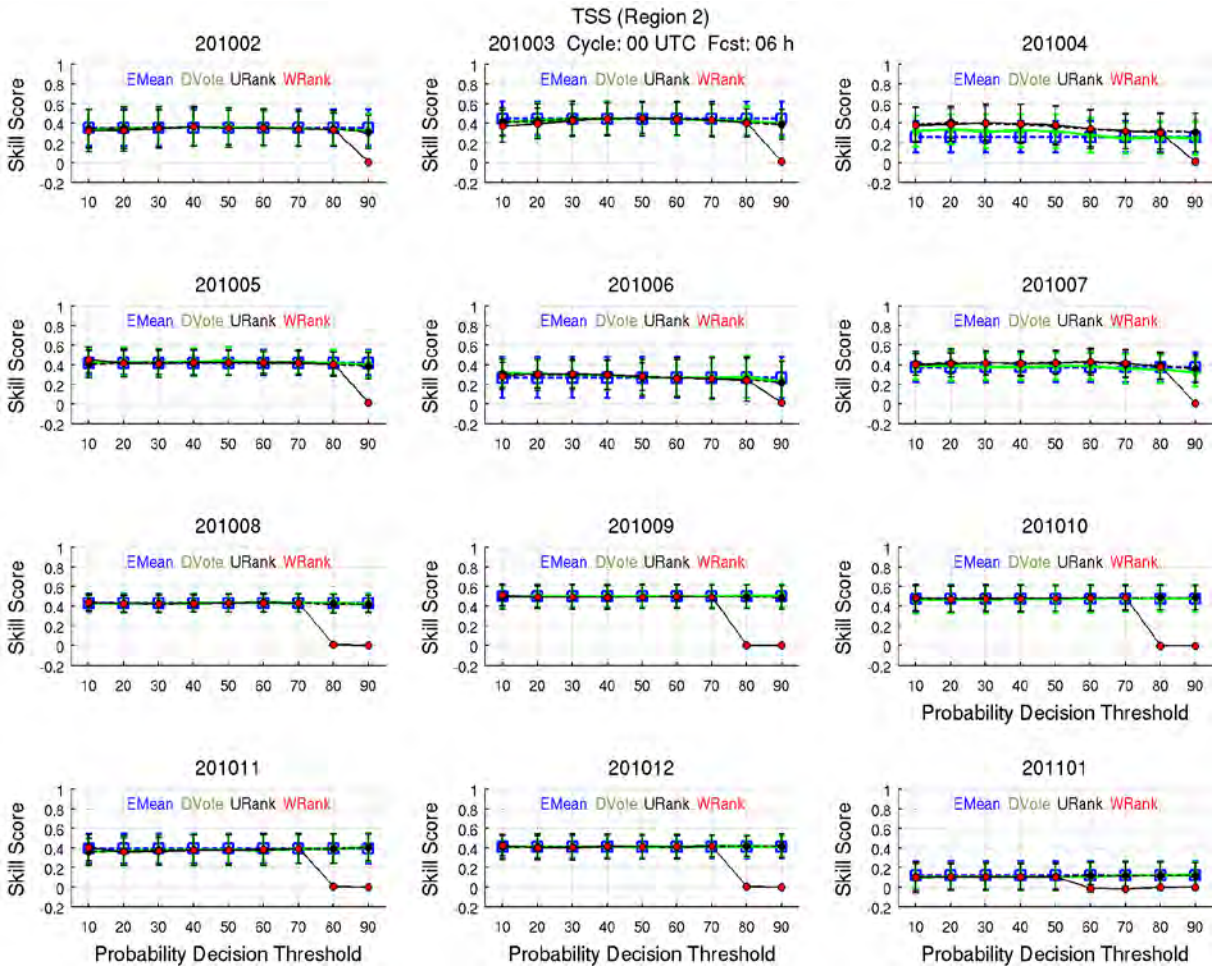


Figure 36. True Skill Score relative to probability decision thresholds (Region 2). Ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) True Skill Scores (left axis) compared to the probability decision threshold (bottom axis). Standard deviations in skill are represented with error bars.

3. Region 3 (Persistently Cloudy)

Figure 37 and Figure 38 display the HSS and TSS results, respectively, for Region 3. The level of skill seen in each ensemble forecast method remains relatively consistent throughout the year. This is consistent with the convective nature of cloud cover in this tropical environment in proximity to the ITCZ. The data we examine points to two primary seasons. The first is February–July when the ITCZ shifts north and the second occurs in August–January when it shifts southward. As before, we examine these seasons utilizing the mean forecast skill to distinguish between ensemble forecast methods with the understanding that the error bars indicate that the differences may not be statistically significant. Because the ensemble behaves deterministically during the latter months we will limit our discussion of skill to February–July.

The HSS shapes are consistent with those we see in Region 1 and Region 2. Here, however, there is more distinction between probability forecast methods. In general, probability forecasts demonstrate less skill than the mean, but the uniform ranks and democratic voting methods perform equally better than the weighted ranks below the 30% decision threshold and above 80% decision threshold. The tendency for the weighted ranks method to forecast clear at lower probabilities in this region produces a significant amount of false alarm.

The TSS, however, indicates that probability forecasts demonstrate more skill at lower decision thresholds. The mean and control forecasts produce more correct rejections, but the overwhelming number of correct rejections by all forecasts makes the false alarm rate much smaller than the hit rate. This makes the ratio between the hits and misses the dominant factor in regions where clouds abound. Therefore, probability forecasts perform better in this region than the mean at lower decision thresholds. The value of correct rejections is no completely discounted as evidenced by the minimal improvement in the TSS over deterministic forecasts.

Returning to the HSS, we see in July that the diminished skill of the probability forecasts is not only limited to lower decision thresholds, but extends to 60%. In this month, probability forecasts produce an inordinate number of false alarms. This brings down the HSS because fewer hits are recorded relative to the number of correct rejections

recorded by the ensemble mean forecast. However, the tradeoff between hits and false alarms and correct rejections and misses results in similar forecast skill as seen in the TSS.

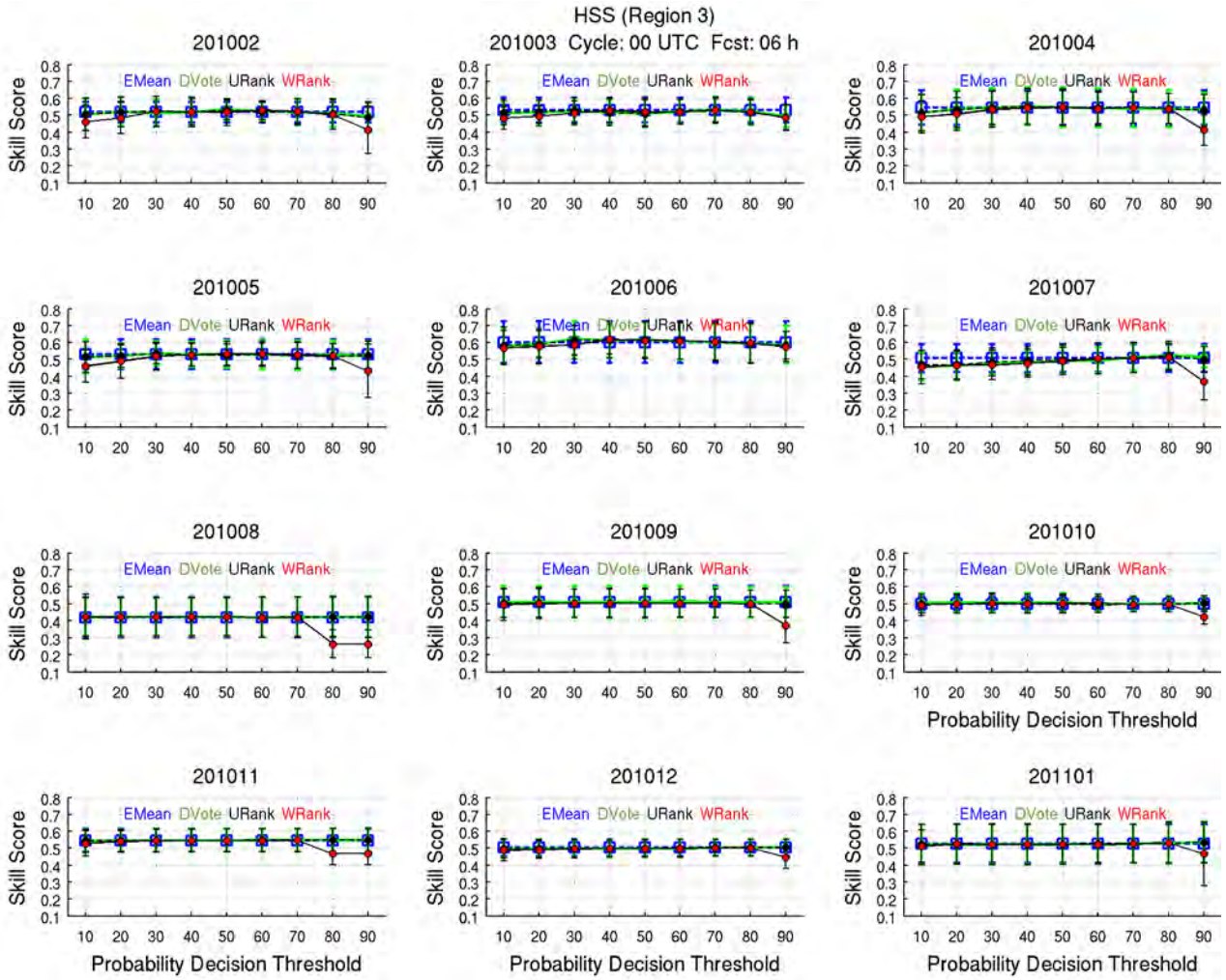


Figure 37. Heidke Skill Score relative to probability decision thresholds (Region 3). Ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) Heidke Skill Scores (left axis) compared to the probability decision threshold (bottom axis). Standard deviations in skill are represented with error bars.

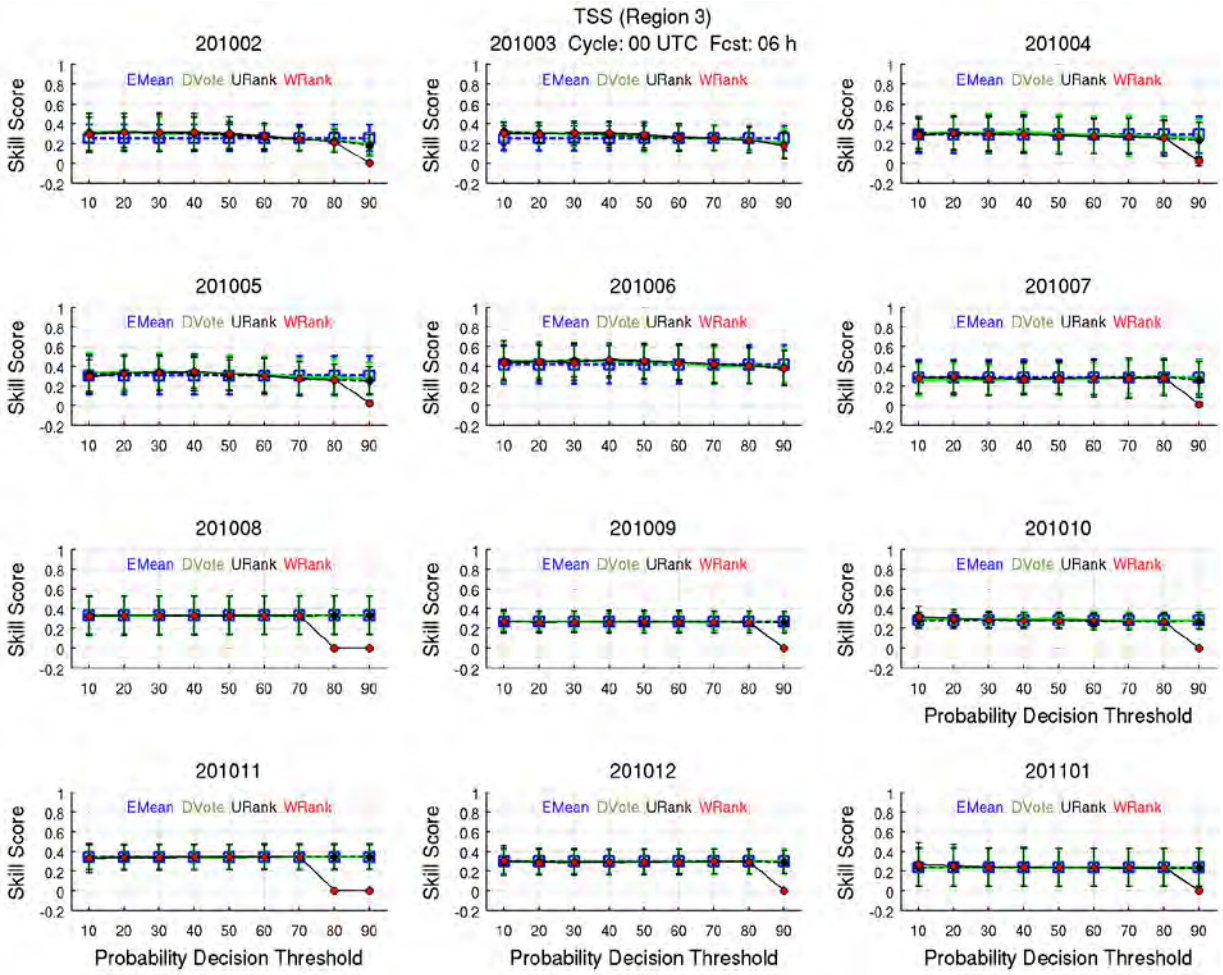


Figure 38. True Skill Score relative to probability decision thresholds (Region 2). Ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) True Skill Scores (left axis) compared to the probability decision threshold (bottom axis). Standard deviations in skill are represented with error bars.

4. Summary

Comparisons of the HSS and TSS reveal important differences. The HSS, a modification of the percent of forecasts correct, should be preferred by users who are indifferent concerning the outcomes of the 2x2 contingency table: hits (misses) are just as desirable (undesirable) as correct rejections (false alarms). We ascribe this behavior to users with a high risk tolerance (10%–30%). These users are willing to accept considerable risk to obtain a clear image. Conversely, the TSS should be preferred by users who are extremely averse to false alarms as compared to missing an opportunity to collect an image. Low tolerant users (70%–90%) are more sensitive to collecting cloudy imagery and therefore, would rather forgo image collection than accept the risk of collecting a cloudy image. Medium tolerant users (40%–60%) may find value in considering both definitions of skill. When the ensemble behaves deterministically from August through January, all users benefit equally except for low tolerance users who use the weighted ranks method.

All forecasts produce significant skill for all users in Region 1. High and medium tolerant users would be better served to use ensemble probability forecasts based on skill score (HSS) results. The probability forecasts demonstrate greater skill from February to June with the most significant differences occurring in March. The high probability of clear conditions along with the tendency for the probability forecasts to predict clear bolsters the number of hits in this region. Low tolerant users experience little difference in forecast skill (TSS) between all forecast methods except the weighted ranks method which tends to under forecast clear with respect to high probability decision thresholds.

The ensemble mean forecast demonstrates better skill in Region 2. Users who use the HSS, because they are not overly concerned with collecting cloudy imagery, find that the mean forecast performs better than all ensemble forecast from February to July. Using the HSS, medium tolerance users experience no difference in skill between forecast methods except in July when the ensemble mean forecast provides the most skill. If the medium user prefers to use the TSS, the ensemble probability forecasts should be preferred to optimize clear verses cloudy collections in the month of April. High tolerance users find the mean ensemble forecast to be the best option based on TSS results.

The ensemble mean forecast is the best option for most users in Region 3. High tolerance users find that the weighted ranks method produces a high number of false alarms and are better off choosing the ensemble mean forecast over other forecast methods. We find this to be the case from February through July. Medium tolerance users employing the HSS are indifferent to the forecast methods except in June (July) when the ensemble probability (mean) forecast performs better. Medium tolerance users employing the TSS should prefer probability forecasts. Although significant differences are not noted in April and July, probability forecasts provide more skill than the mean during the first six months of the dataset. Low tolerance users gain the most skill in the employment of the ensemble mean forecast.

One of the key challenges with the use of the weighted ranks method is that it favors clear conditions at lower probability decision thresholds and cloudy at higher thresholds. This arises from the method in which we calculate the probabilities. The probabilities are taken from the cumulative probability distribution function of the verification rank histograms that corresponds to the democratic voting probability of the ensemble members.

Figure 39 is provided to help visualize the difference in probability after the rank histogram conversion (bin 21 not included). The probability distribution function (top) causes the weighted ranks method to prefer clear at lower probability decision thresholds and prefer cloudy at higher probability thresholds. The dashed line marks the point where weighted ranks probability matches the democratic voting probability. The February HSS chart from Region 2 (lower) is extracted from Figure 35. This chart is chosen at random. The HSS is used instead of the TSS because clear and cloudy tendencies can easily be inferred from the proportion calculation.

The dashed blue line indicates that the democratic voting (blue) and the weighted ranks (red) histograms intersect at ~39% probability of clear. Below the line, the weighted ranks method shifts the ensemble forecast towards clear. Above the line, the ensemble forecast is shifted towards cloudy.

Consider the most extreme users. The user who only needs two ensemble members (10%) to forecast clear conditions before collections are taken. Based on the

weighted ranks conversion, it only takes one ensemble member to forecast clear before the probability (15%) exceeds the threshold and the recommendation to collect is given. At the opposite end, the low tolerant user (90%) needs at least 18 of the ensemble members to forecast clear before a collection is taken. As seen in Figure 39, when all ensemble members forecast clear, the probability forecast is less-than 90%. Therefore, the possibility exists that the weighted ranks method may not forecast 90% probability of clear even with the fractional probability received from bin 21.

High probabilities of clear, in this case, may not be reached because the cumulative probability shifts the probability of clear towards lower probabilities. This occurs when the verification rank histogram suggests a strong cloudy bias – bin 1 of the under-dispersed verification rank histogram is larger than bin *n* (the last bin) because observations are most often drier than the ensemble. The 100% probability of clear (bin 21) is not shown in the figure. Instead, the red bar at 100% is the minimum probability of clear when all of the ensemble members are below the threshold.

This minimum probability of clear is adjusted relative to the forecast. If the minimum ensemble forecast value below the threshold is 0%, then the probability contained in bin *n* is added to the forecast probability to obtain 100% probability of clear. If the minimum ensemble member's forecast value is 30%, the probability contained in the last verification bin is not used. A fraction of the value is added for all other cases based on the method discussed in section 4.

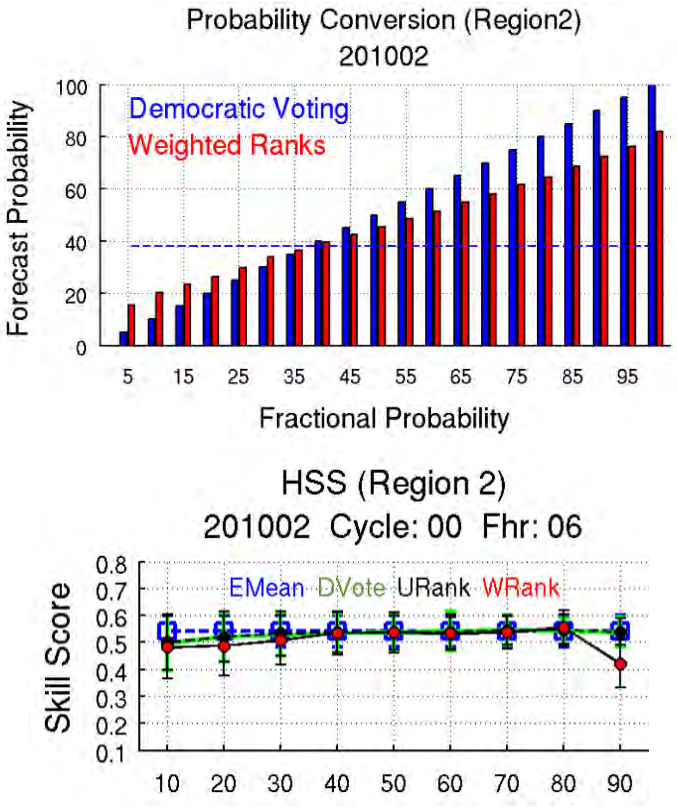


Figure 39. Probability conversion chart. Probability of clear is higher with the weighted ranks method than the democratic voting method below 40% probability of clear (top). Thus, the ensemble over forecasts clear and has less skill than the democratic method below 40% probability of clear (bottom). The weighted ranks method rarely forecasts above 90% probability of clear (top). Hence, a reduction is skill above 90% (bottom).

The democratic voting and uniform ranks methods also have a tendency to forecast clear (cloudy) conditions too often at lower (higher) probabilities, but for different reasons than the weighted ranks method. A high tendency for clear (cloudy) predictions by these forecasts suggests that the certainty, as defined by the ensemble, is rarely greater than 90%; the democratic voting method has two or more outlier forecasts. The fact that the ensemble favors clear forecasts at lower decision thresholds and experiences a higher number of false alarms is not surprising. It is quite possible for the mean of the ensemble to be greater than the 30% cloud-free threshold when 2/20 member forecast are below the threshold.

However, the ensemble has an increased number of misses at higher probability decision thresholds. Although the ensemble frequently forecasts clear conditions, the

probability of clear conditions tends to be below 90%. Since the driest value between WWMCA and GEFS is used to initialize each member forecast the ensemble should have a dry bias. This is either an initialization or advection problem.

There are three possible effects associated with the method in which we initialize GACE with WWMCA, two of which have implications on the ensemble spread. 1) WWMCA CPS values fall above all ensemble members and the ensemble spread is not modified. 2) WWMCA CPS values lie below all ensemble members and the ensemble spread is reduced to zero. 3) WWMCA CPS values lie between the CPS values of two ensemble members and the ensemble spread is reduced to the standard deviation between the lowest member and the WWMCA value. Effect 1 has no forecast implications other than the fact that WWMCA is not used to initialize the model. Effect 2 returns the ensemble initial cloud field to an unperturbed state. Effect 3 reduces the characterization of uncertainty in the initial cloud field.

Which of these effects could be responsible for the ensemble's tendency to avoid forecasting 100% clear conditions? In order for Effects 1 to be the cause, WWMCA would have to be above the 30% threshold and GEFS would have to straddle the threshold. If Effect 2 is the case, the WWMCA value is above the threshold and the ensemble has a moist biased spread. Effect 3 requires that the WWMCA value be above the threshold and GEFS have at least one member forecast below the threshold and one above the WWMCA value. These effects are plausible, but advection can also attribute to the cloudy tendency.

The consumers of Air Force Weather information rarely have the luxury of waiting for perfect weather. Therefore, diminished skill at higher probability decision thresholds is not a major concern. Of more interest is the skill at lower probability thresholds. It is at these thresholds that we believe military decision makers reside. Operations can be very sensitive to cloud filled imagery; however, they also are willing to take significant risks to collect high priority imagery. Upon this premise, we elect to use 20% probability decision threshold as we move further in our discussion of ensemble skill in cloud-free forecasting.

In addition, we relax our strict stipulation between high, medium, and low users. Although Air Force operations are driven by objective completion, they are not without cost consideration. The military must also operate efficiently, so regularly collecting cloudy imagery is not desired. Therefore, we examine ensemble skill using both the HSS and TSS at the 20% probability threshold in the sections to come.

2. CYCLE AND HOUR VARIATIONS

Figure 40 and Figure 41 show the annual skill of the ensemble forecasts at the 0000 UTC and 1200 UTC cycles over the 24-h forecast period for Region 1, Region 2 and Region 3. The annual mean skill score and mean standard deviation are calculated with respect to cycle, forecast hour, and region. Again we see, as with the probability decision threshold charts, the error bars overlap when skill is combined by cycle and hour. Understanding that errors can quite possibly exist in our interpretations, we continue to use the mean skill value to compare the ensemble forecast methods. Because there are no significant annual differences in TSS between the ensemble forecast methods, we omit this measure of skill in our cycle and hourly evaluation. We evaluate each region in sequence.

In Region 1, slight differences are observed in skill with cycle and forecast hour. First, the HSS suggests that the skill depreciation of the ensemble forecasts is gradual with the 0000 UTC cycle. In contrast, we see a significant drop in skill in the 12-h forecast of the 1200 UTC cycle. By the 18-h forecast, forecast skill levels off. After 24 h the skill matches that of the 0000 UTC cycle forecast for all ensemble forecast methods. We also note that probability forecasts tend to perform better in this region with both forecast cycles. This arises in large part because the probability forecasts have a clear tendency at 20% probability of clear. Annually, these forecasts produce more hits than the mean ensemble forecast.

The hourly differences in skill of the ensemble forecasts could possibly be attributed to cloud cover differences in the day and night cloud characterizations of WWMCA and ensemble initialization. Comparing the skill difference between cycles and the standard time zone conversions in Table 9, we see a distinct correlation between forecast skill and time of forecast initialization. The 0000 UTC forecast is initialized

before sun rise when visible imagery is not available. However, it is verified during the day when visible imagery is included in the WWMCA analysis. In contrast, the 1200 UTC cycle is initialized during the day, but the first two forecast hours are verified at night. In the case of the 1200 UTC cycle, clouds that were characterized with visible imagery in the 1200 UTC cycle appear to be missed in the WWMCA cloud cover depiction at night. It follows that, forecasts that favor clear conditions benefit erroneously from the degraded analysis when the clear conditions prevail.

Table 9. Standard time zone conversions. Values represent the mean time zone of each region. Shaded boxes represent nighttime initialization or verification.

Forecast Hour	Region 1 (+4)		Region 2 (+8)		Region 3 (-4)	
	0000 UTC	1200 UTC	0000 UTC	1200 UTC	0000 UTC	1200 UTC
0	4	16	8	20	20	8
6	10	22	14	2	2	14
12	16	4	20	8	8	20
18	22	10	2	14	14	2
24	4	16	8	20	20	8

In Region 2, all ensemble forecasts demonstrate diminishing skill from the 6 h to the 24-h forecast. We also see gradual decline of skill at 0000 UTC but a steeper down-slope of skill with the 1200 UTC cycle, which levels of by the 18-h forecast. In this region, the tendency for the probability forecasts to forecast clear causes more false alarms. The disparity increases with time between ensemble probability and mean forecasts with the 0000 UTC forecast. We also see separation between the ensemble probability forecasts with the 0000 UTC cycle with the weighted ranks method becoming increasingly worse with time. The difference in the forecasts over time, with respect to the 1200 UTC cycle, changes little.

Initialization and verification with WWMCA is less of a problem in this region. The 0000 UTC cycle is initialized during sunlight hours. The gradual decline in skill in the 0000 UTC cycle does not appear to be affected by the lack of visible imagery at the 12 h and 18-h forecast times. Furthermore, the 1200 UTC cycle initializes at night, and we do not see indications of the clear bias in the probability forecasts.

In Region 3, the skill of the ensemble forecasts decrease over time with the 0000 UTC cycle but increases with the 1200 UTC cycle. Looking at the result of the 0000 UTC cycle, we see that the ensemble mean forecast performs better than probability forecasts. As forecast lead-time increases, the disparity between the weighted ranks method and other probability forecasts increase. After 1800 UTC, however, ensemble skill rebounds. The mean performs better than probability forecasts but to a lesser degree at the 1200 UTC cycle where the skill improves with forecast lead-time.

The 0000 UTC forecast initializes at night and the 1200 UTC forecast initializes during the day. It appears that the night time initialization causes problems for the probability forecast methods, which routinely over forecast clear conditions. The analysis, without the benefit of visible imagery, can underestimate the amount or presence of clouds. This causes the forecasts to underestimate cloud cover at future time-steps. When additional cloud cover is introduced with the use of optical sensors, the analysis tends to be cloudier than the forecast. Therefore, forecasts that favor clear are less skillful.

The differences in skill of the ensemble forecasts at 20% probability decision threshold are not significant but provide information about the role of WWMCA in the verification process. Skill differences suggest that WWMCA can detract or bolster forecast skill. If the initialization occurs at night, and clouds are not detected, verification performed during the day with the benefit of visible imagery can favor forecasts with a cloudy bias. If the initialization occurs during the day, the reduction in cloud detection by WWMCA could be beneficial to forecasts that preference clear. We are careful to note, however, that these signatures were not seen in Region 2 because of the variability in cloud cover frequency.

Cloud conditions in Regions 1 and 3 are persistent. Therefore, the fact that probability forecasts perform better in Region 1 and the ensemble mean performs better in Region 3 is not surprising. We conclude that variations in WWMCA day and night cloud cover amounts only attribute to the magnitude of the skill differences and are not the cause of the differences themselves. With this in mind, we proceed with ensemble

comparisons. Furthermore, the differences between forecast cycles and hours are not statistically significant, so we limit our future evaluations of ensemble skill to the 0000 UTC cycle and 6-h forecasts.

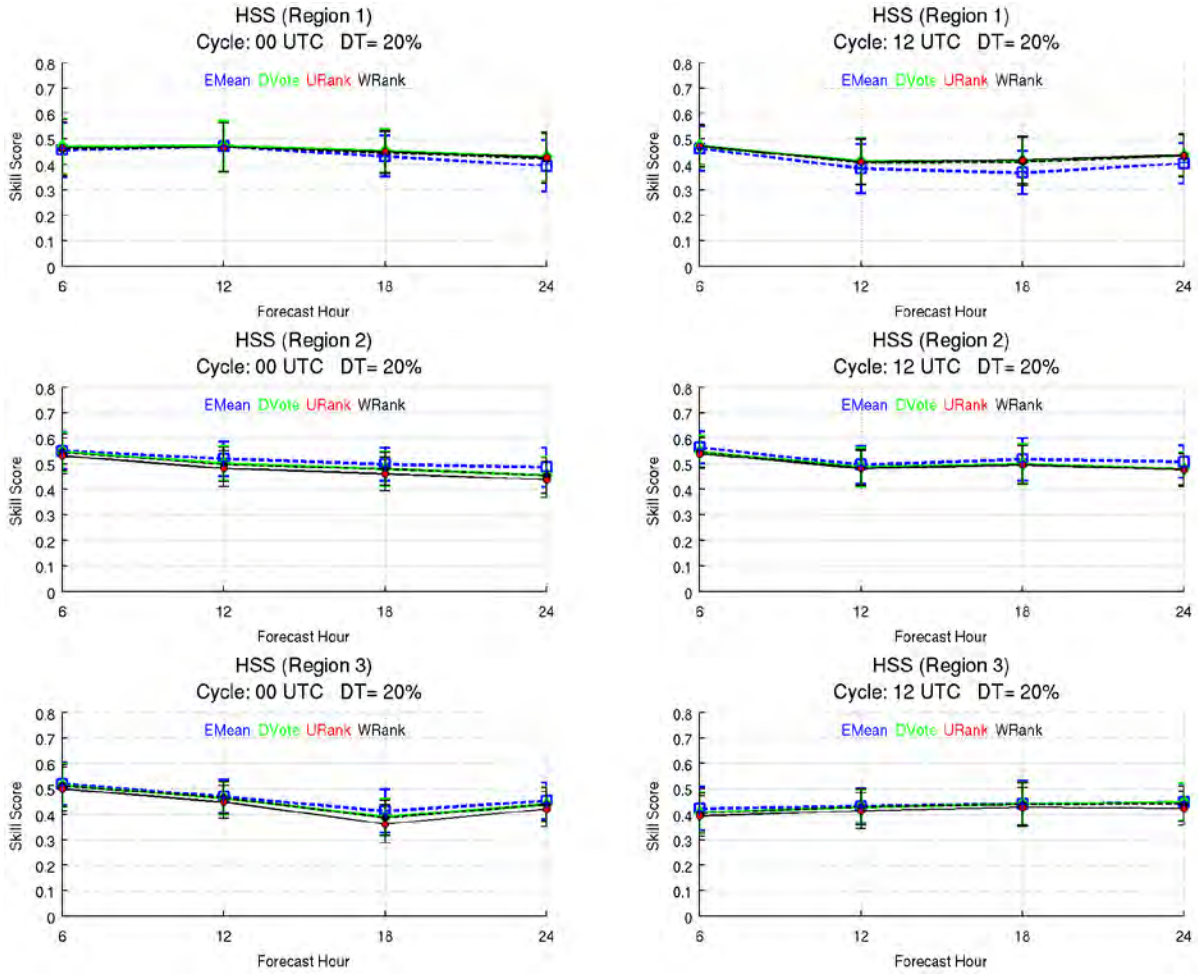


Figure 40. Heidke Skill Score variations with cycle and hour for Regions 1, 2, and 3. Ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) Heidke Skill Scores (left axis) compared hourly (bottom axis). Standard deviations in skill are represented with error bars.

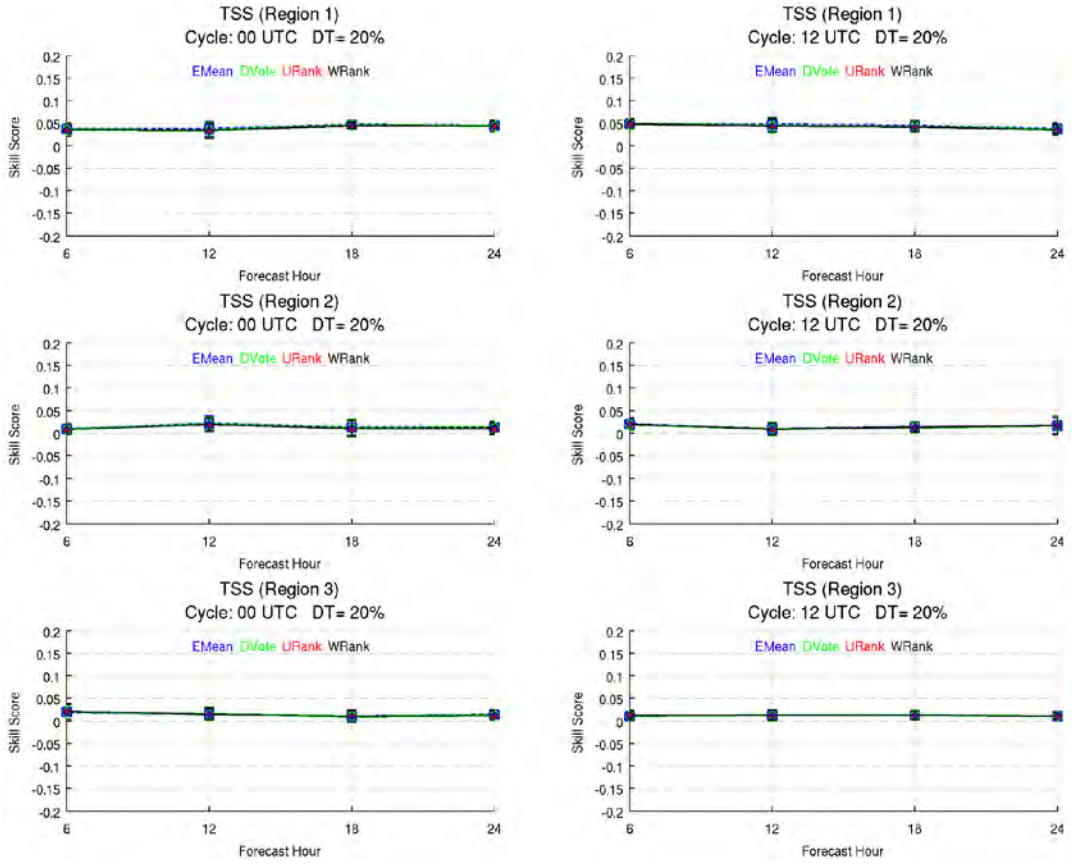


Figure 41. True Skill Score variations with cycle and hour for Regions 1, 2, and 3. Ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) True Skill Scores (left axis) compared hourly (bottom axis). Standard deviations in skill are represented with error bars.

3. FREQUENCY VARIATIONS

We begin our analysis of ensemble skill by evaluating ROC diagrams for each region. The value of HT, MS, CR, and FA are totaled regionally, and then ROC diagrams are constructed using the hit rate H and false alarm rate F calculated at each probability decision threshold (0%–100%). Users with a high tolerance for cloudy imagery will accept more risk and attempt to collect imagery more often. Less tolerant users are risk-averse and choose to wait until the chance of collecting clear imagery is in their favor. Therefore, the number of hits and false alarms tend to have an inverse relationship with probability decision threshold while correct rejections and misses maintain a direct relationship.

Plotting the hit rate H versus the false alarm rate F to produce ROC curves, reveals information about the skill of the forecast relative to a user's decision threshold. When H > F, the forecast is considered to have skill. No skill or random forecasts are characterized by a hit rate that matches the false alarm rate. In general, real forecasts do not fall below the random skill line (H = F), but remain between perfect skill (1) and random skill (.5) (Wilks 2001).

Although we find our results to be consistent with real forecasts, the ROC diagrams in August 2010 through January 2011 show that the probability forecasts have ROC curves characteristic of deterministic forecasts; the skill of a deterministic forecast does not change with probability thresholds. Extremely under-dispersed ensembles exhibit this characteristic when plotted on a ROC diagram. Because this phenomenon is seen across each region and forecast hour, we are confident that this behavior is not an indicator of increased certainty in cloud cover development and/or advection.

After carefully examining the algorithms to ensure that all months are calculated in the same manner, the lack of diversity in member forecasts point to assimilation and/or model changes performed by NCEP in the fall of 2010 (Zhu et. al. 2011). It appears that a reduction in the ensemble spread designed to refine the extended forecast reduced the utility of the ensemble in our short term forecast application. Therefore, model performance during the latter months of our dataset is included for completeness but is not discussed in great detail.

After using ROC diagrams to evaluate the ensemble skill in producing cloud-free forecasts in each region, we turn to the HSS to evaluate how much of the forecast skill can be attributed to ensemble versus random chance. The TSS is used to evaluate how well the ensemble distinguishes between clear and cloudy events. The ensemble's performance is evaluated relative to the monthly averaged probability of clear conditions within each region.

The monthly mean HSS and TSS are calculated as follows. The skill score at each grid point is placed in 1 of 11 bins depending on the grid point's monthly cloud cover frequency (calculated from daily WWMCA values). Grid points with cloud cover frequency of 0 are placed in bin 1. Frequencies > 0 and < 10 are placed in bin 2; frequencies > 10 and < 20 are placed in bin 3…and frequencies > 90 and < 100 are placed in bin 11. It is also useful to note that skill at p(clear)=100% represents the skill at p(clear)>90%. The average HSS of each bin is calculated and plotted in Figure 43. The number of forecasts evaluated at each frequency is also plotted (thin, blue, dashed line). Error bars are not included to improve the readability of the charts. However, they are of the same magnitude of those seen in our decision threshold and cycle evaluations.

Correct rejections are the dominant contributor of skill in clear areas, and hits are the dominant contributors in cloudy areas. The HSS primarily evaluates the ability of forecasts to correctly detect hits and correct rejections. The TSS, which fosters a distinction between positive and negative outcomes, is used to evaluate the overall skill of the ensemble. However, evaluating the elements of the 2x2 contingency table directly is also useful.

1. Region 1 (Persistently Clear)

a. ROC Diagram

The ensemble demonstrates skill for each month. Variations in skill between each ensemble forecast (mean, democratic voting, uniform ranks, and weighted ranks) coincide with WWMCA seasonal cloud cover changes identified in Figure 25. From February to March, winter frontal systems—observed every 3–5 days—become less frequent causing a decrease in the mean cloud cover. During this transition, the ensemble forecasts demonstrate a deterioration of skill. With fewer cloudy grid points in

March, the ensemble mean produces correct rejections at the same rate as misses. This reduction (increase) in the number of correct rejections (misses) increases F (decreases H). This is the first indication that the mean forecast has a cloudy preference during this season. Conversely, probability forecasts produce more false alarms than misses at probability decision thresholds less than 50%. The ratio reverses above 50%.

In spring (April–June) frontal systems move through the region every 5–7 days and the number of cloudy grid points continues to decrease. The majority of the grid points are clear at least 60% of the time. However, the ROC curves do not indicate a significant shift to the left of the diagram. As the opportunities for clear increases, we expect the number of hits to increase and false alarms to decrease relative to previous months. This indicates that the ensemble has challenges identifying the occasional cloudy event.

From June to August, the southwest quadrant of Region 1 sees an increase in cloud cover during the South Asian Monsoon. The forecasts correctly reject the monsoonal cloud cover and the false alarm rate decreases, and the ROC area increases. By September, the prevailing clear condition makes outcomes other than hits less likely.

b. Heidke and True Skill Score

From February through April, the HSS (Figure 43) indicates that the ensemble forecasts generally demonstrate skill above 20% probability of clear conditions—grid points with less than 20% probability of being clear rarely exist in this region. As with the ROC curves, we see a decrease in the HSS. We also note that the disparity between the ensemble mean and probability forecasts is most prominent between 40% and 70% probability of clear, where uncertainty is highest. The skill of the control forecast is consistently found between the ensemble uniform ranks and mean methods.

Comparing the TSS (Figure 44) values, which explicitly include incorrect forecasts, we see significant differences between 40% and 90% probability of clear. The ensemble mean appears to have a cloudy propensity in that it consistently produces more correct rejections and misses than probability forecasts above 40% probability of clear.

Although the skill of the control forecast is most often equitable to the ensemble mean, it never performs better when conditions are sufficiently sampled.

The ensemble mean and control forecasts produce a relatively high number of misses when the probability of clear conditions is high as compared to probabilistic forecasts, but the number of misses becomes insignificant in calculating the hit rate because the hit occurrence is so large. Therefore, the hit rate of the ensemble mean and control forecasts remain close to 1. In like manner, the preference for cloudy conditions increases the number of correct rejections and reduces the number of false alarms, which intern reduces the false alarm rate. The increase in the hit rate and decrease in the false alarm rate results in an increase in the TSS of the ensemble mean and control forecasts relative to the probability forecasts.

In May and June, HSS differences between ensemble forecast methods can be identified even though cloudy conditions are relatively rare. These differences are an indication of the ensemble's ability to distinguish between clear and cloudy events when faced with the threat of cloudy conditions. In areas where the sample size is sufficiently large, the HSS provides an indication of which forecasts most often deviates incorrectly from a clear prediction. The tendency of the ensemble probability forecasts to predict clear conditions at this decision threshold fosters improved skill over the ensemble mean and control forecasts, which have a tendency to forecast cloudy conditions even when the probability of clear is high.

The overwhelming number of hits produced by all forecasts in this region has significant implication on the TSS of the forecast methods. Outcomes other than hits only amount to ~3% of the forecast verifications. In this environment, misses become inconsequential in calculating the TSS, and skill differences rely less on the magnitude of hit rate and more on the false alarm rate. Moreover, TSS comparisons reveal the ability of the ensemble to minimize false alarms and maximize correct rejections. The large number of hits masks the impact of missing an opportunity to collect a clear image.

The ensemble mean and control appear to demonstrate more skill, per the TSS, in May than in June at high probabilities of clear. Although more misses are recorded by these forecasts than probability forecasts, the ensemble mean and control

forecasts produce as much as 20% more skill in the month of May. The skill here is not demonstrated in the number of misses avoided; it too arises from the ensemble's ability to limit false alarms and correctly predict cloudy events. At 80% probability of clear conditions, probability forecasts produced more false alarms than misses. This suggests a strong clear bias.

The ensemble mean and control are only slightly better than the probability forecasts in June. Probability forecasts continue to produce false alarms at a higher rate than misses. Although this is not the case with the ensemble mean and control forecasts, the ratio of false alarms to misses is close to 1.

In July, the HSS and TSS tell similar stories about the skill of the ensemble mean and control as compared to probability forecasts. Based on the number of days clear conditions are observed, several grid points experience intermittent cloud cover. The control and mean forecasts perform better than probability forecasts when clear conditions are observed less-than 50% of the time. In addition, the performance of the democratic voting method has noticeable improvement over other probability forecasts at 10% and 50% probability of clear.

The tendency for probability forecasts to favor clear conditions causes degradation in HSS below 50% probability of clear and improvement in skill above 50% probability of clear. Below 50% probability of clear, probability forecasts produce more false alarms than misses. Above 50% probability of clear, the mean and control forecasts produces far more misses than probability forecasts. Therefore, probability forecasts perform better above 50% probability of clear, and the ensemble mean and control perform better below the 50%.

Biases identified in section 1 of this chapter are evident in our evaluation of skill in Region 1. The natural assumption is that forecasts that have a clear tendency perform better in regions that are predominately clear. We find, however, that this assumption does not always hold true. These forecasts must also be able to identify changes in the predominant cloud cover. If not, an increase in the number of false alarms will ensue. We find that the ensemble mean and control forecasts perform best in most

cases. In situations where probability forecasts perform better, the differences in skill are relatively small except for July where the frequency of clear conditions is not concentrated at 80%–90%.

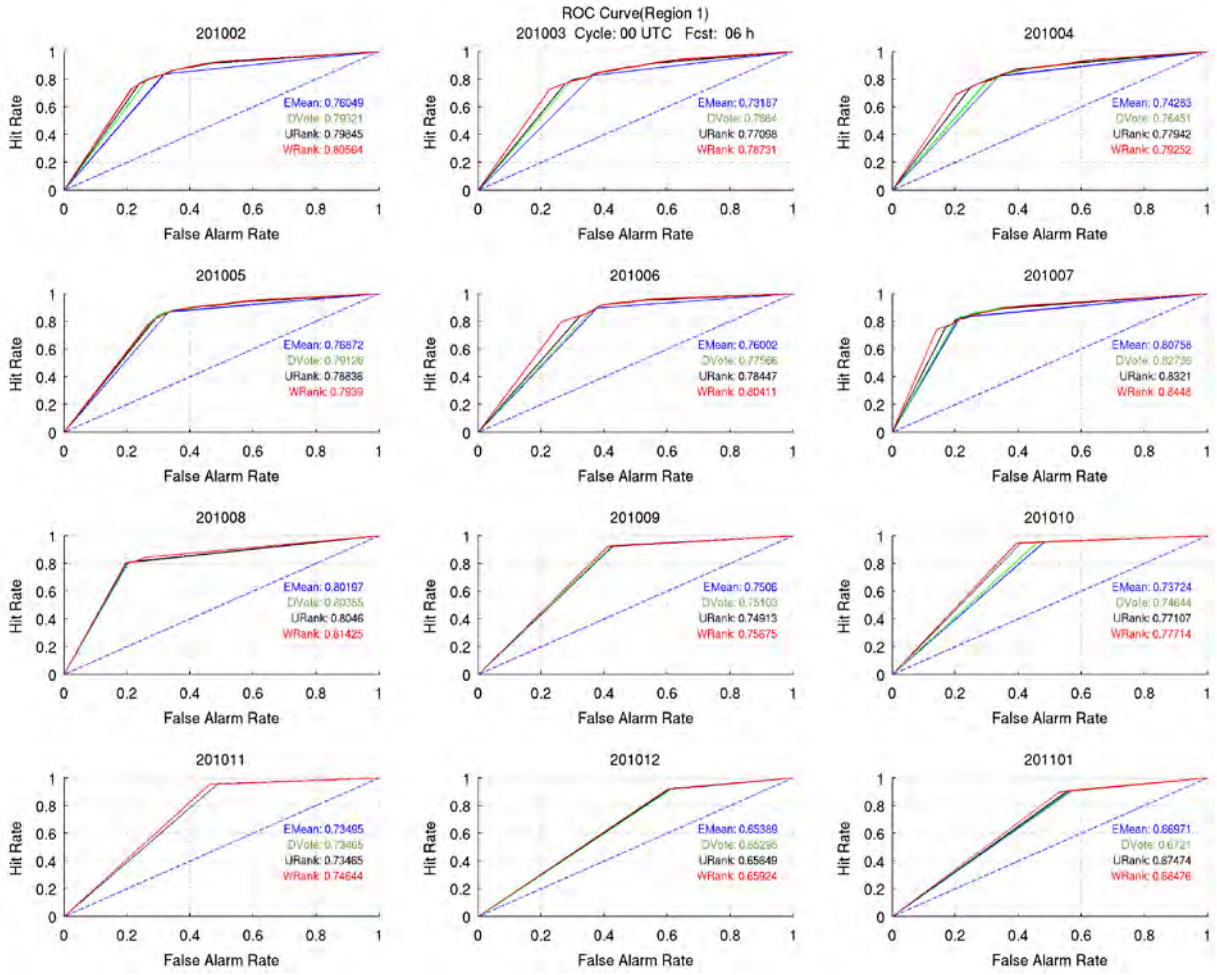


Figure 42. ROC diagrams for Region 1. ROC diagram for 6-h forecast and 0000 UTC cycle. The hit rate (left axis) and false alarm rate (bottom axis) are plotted for the ensemble mean, democratic voting, uniform ranks, and weighted ranks method, and the their ROC areas are included for numerical comparisons.

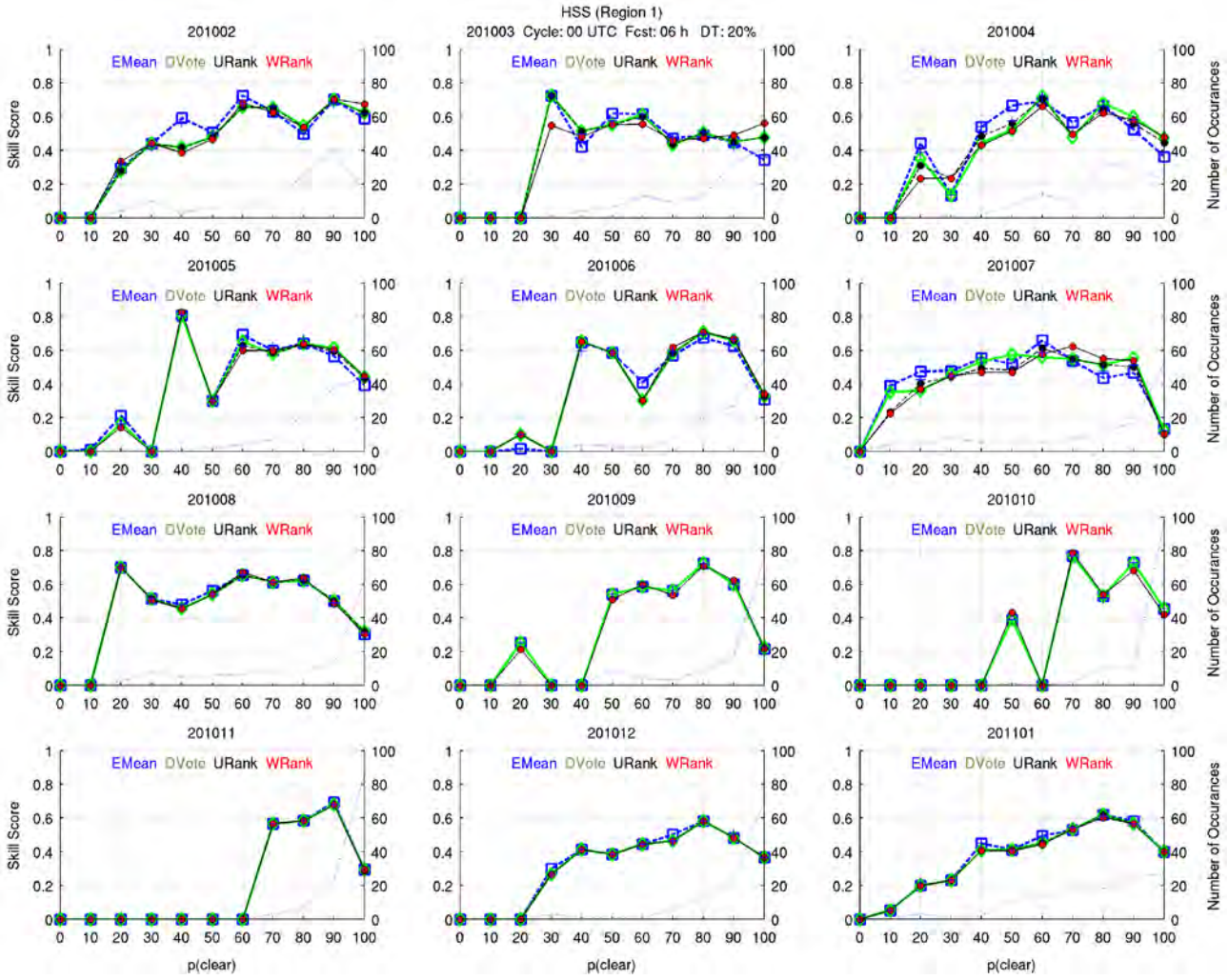


Figure 43. Heidke Skill Score relative to probability of clear conditions (Region 1). Ensemble Heidke Skill Score (left axis) compared to the probability of clear conditions (bottom axis). Thin dashed line indicates the number of grid points (right axis) evaluated for each *p(clear)* threshold (bottom axis).

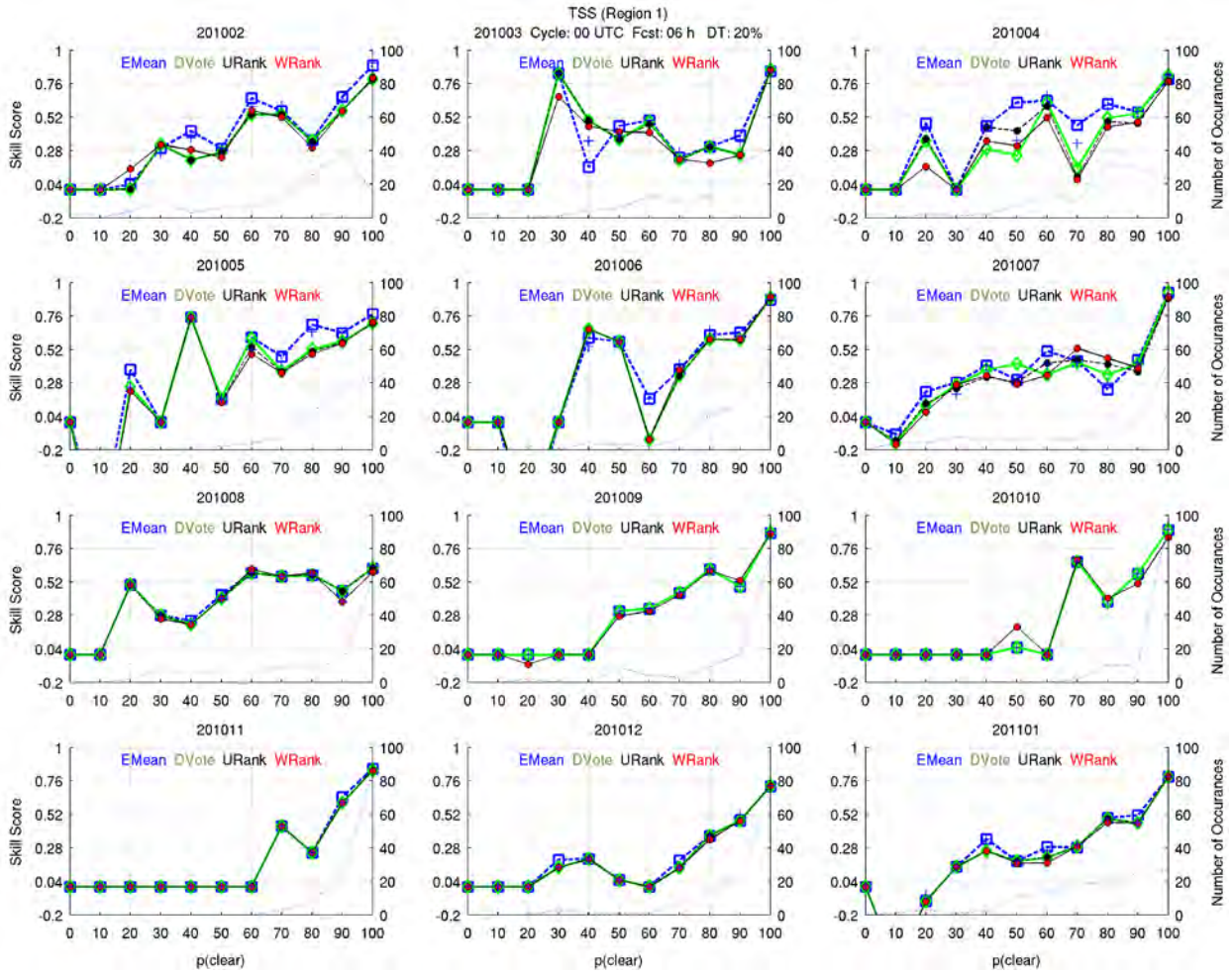


Figure 44. True Skill Score relative to probability of clear conditions (Region 1). Ensemble True Skill Score (left axis) compared to the probability of clear conditions (bottom axis). Thin dashed line indicates the number of grid points (right axis) evaluated for each *p(clear)* threshold (bottom axis).

2. Region 2 (Variable Cloud Cover)

a. ROC Diagram

Region 2 ROC diagrams are shown in Figure 45. Ensemble probability forecasts demonstrate more skill than the ensemble mean during each month. From February to April, the disparity in skill between the ensemble probability and mean forecast increases as clear opportunities decrease. Distinctions in skill between ensemble probability methods also become apparent in April. These distinctions are maintained through July during the influx of monsoonal cloud cover, which peaks in June (Figure 27). From August to January, the skill of all ensemble forecast methods becomes almost indistinguishable. The ROC curves suggest that if we use one forecast method for all users, probability forecasts will provide more skill to the accumulation of user than the mean forecast. What about our specific user who has a decision threshold of 20%?

b. Heidke and True Skill Score

We find that HSS differences between the ensemble forecasts are highly sensitive to cloud cover frequency in Region 2 (Figure 46). We also note that 50% cloud cover frequency often marks a distinctive shift in performance of the ensemble forecasts. Some months show more variability in skill than others, but all ensemble forecasts produce significant skill from 20%–90% cloud cover when the sample size is sufficiently large. The February and April spikes in skill at 10% and 90%, respectively, are attributed to an insufficient sample size.

We first examine the HSS of the ensemble forecasts at probabilities of clear below 50%. The ensemble mean and control forecasts perform better than probability forecasts. The skill of the control forecast only differs from the ensemble mean at frequencies with small sample sizes. The weighted ranks method performs worse than other probability forecast methods. Differences are not apparent between the democratic voting and uniform ranks methods.

The ensemble mean and control forecasts tend to forecast cloudy more often than probability forecasts. Therefore, the differences in HSS arise from the mean and control producing more correct rejections and fewer false alarms than the probability forecasts. Although the democratic voting and uniform ranks methods perform better

than the weighted ranks method, it appears that the increased probability of clear of the uniform ranks method, which arises from the probability of clear occurring between ensemble members, has no identifiable impact on the HSS in February and March. Never-the-less, all probability forecasts demonstrate a tendency toward clear conditions at probabilities of clear below 50%.

These tendencies are also seen in the TSS (Figure 47). The differences between the ensemble mean/control and probability forecasts are not as dramatic as with the HSS. The large number of false alarms produced by the probability forecasts is less of a factor with the TSS because the large number of correct cloudy forecasts and the relatively few opportunities to collect clear imagery (few misses) make the false alarm rate much smaller than the hit rate.

At probability of clear conditions above 50%, probability forecasts perform better than the mean and control forecasts. It is difficult, however, to establish which probability forecast method demonstrates superior skill. The small sample size at higher frequencies and the cloud cover variability within the region makes it difficult to find trends or consistencies in the HSS and TSS. Determining which forecast has the most skill depends on the frequency of clear condition being examined.

April begins the monsoon season when the Indian Ocean monsoon moves north over Eastern China. During this period, the democratic voting method produces the best HSS at all clear condition frequencies. Here we see that the tendency for other probability forecasts to predict a higher probability of clear results in more false alarms when the probability of clear conditions is less than 40%. Above 40%, the differences between probability forecasts are not as significant; but all probability forecasts receive a higher HSS than the ensemble mean and control forecasts.

In contrast to the HSS in April, the TSS suggests that the weighted ranks method produces the most skill below 60% probability of clear conditions. As seen before, the number of correct rejections produced in a predominantly cloudy region moderates the number of false alarms produced by forecasts that favor clear conditions. Therefore, we see a hierarchy of skill from the least likely to forecast clear (EMean) to

the most likely to forecast clear (WRank). The sample size above 60% probability of clear is too small to make sound conclusion, but the mean appears to handle the rare events best at these frequencies.

From May through July cloud cover increases with the onset of the monsoon season. Differences in the HSS of the democratic voting and uniform rank methods are again indistinguishable and compatible with other forecast methods. Noteworthy differences in skill occur at 20% probability of clear. At this probability, the ensemble mean performs best in May and July, and the weighted ranks method performs better in June.

In this monsoon environment where grid points tend to be persistently cloudy, false alarms are insignificant compared to the number of correct rejections and affect little change on the false alarm rate. Hence, TSS comparisons become measures of a forecast's ability to increase hits and reduce missed opportunities. The tendency for the mean ensemble forecast to predict cloudy conditions produces more correct rejections, but probability forecasts demonstrate more skill in that they produce more hits at probabilities below 50%. The mean ensemble forecast demonstrates superior skill when the probability of clear is above 60%, but the sample sizes at these frequencies are too small to adequately judge the results.

In Region 2, we find that the HSS and TSS articulate different conclusions about forecast skill. The HSS suggests that probability forecasts tend to perform better, because more hits are achieved in the midst of cloudy and highly variable conditions. Conversely, the TSS suggests that the mean and control forecasts may perform better when probability forecasts produce false alarms at a rate comparable to the number of hits. The decision maker, who has a decision threshold of 20% and a high priority for collecting clear imagery, should use the TSS and preference the weighted ranks method, which performs best the majority of the first six months. If the decision maker is also concerned with collecting too many cloud filled images, the TSS is consulted and the ensemble mean forecast is used, which performs best for the majority of the first six months.

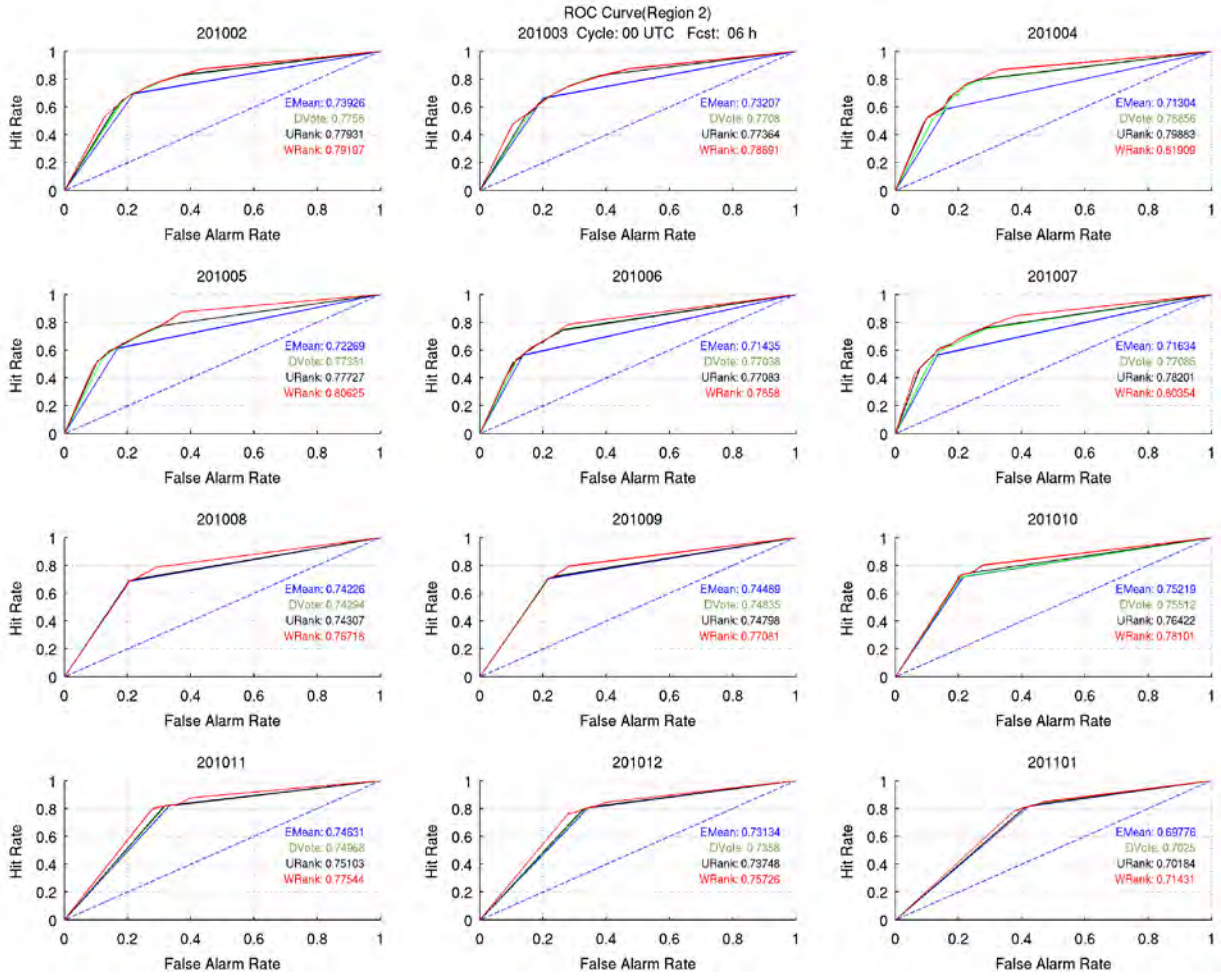


Figure 45. ROC diagrams for Region 2. ROC diagram for 6-h forecast and 0000 UTC cycle. The hit rate (left axis) and false alarm rate (bottom axis) are plotted for the ensemble mean, democratic voting, uniform ranks, and weighted ranks method, and the their ROC areas are included for numerical comparisons.

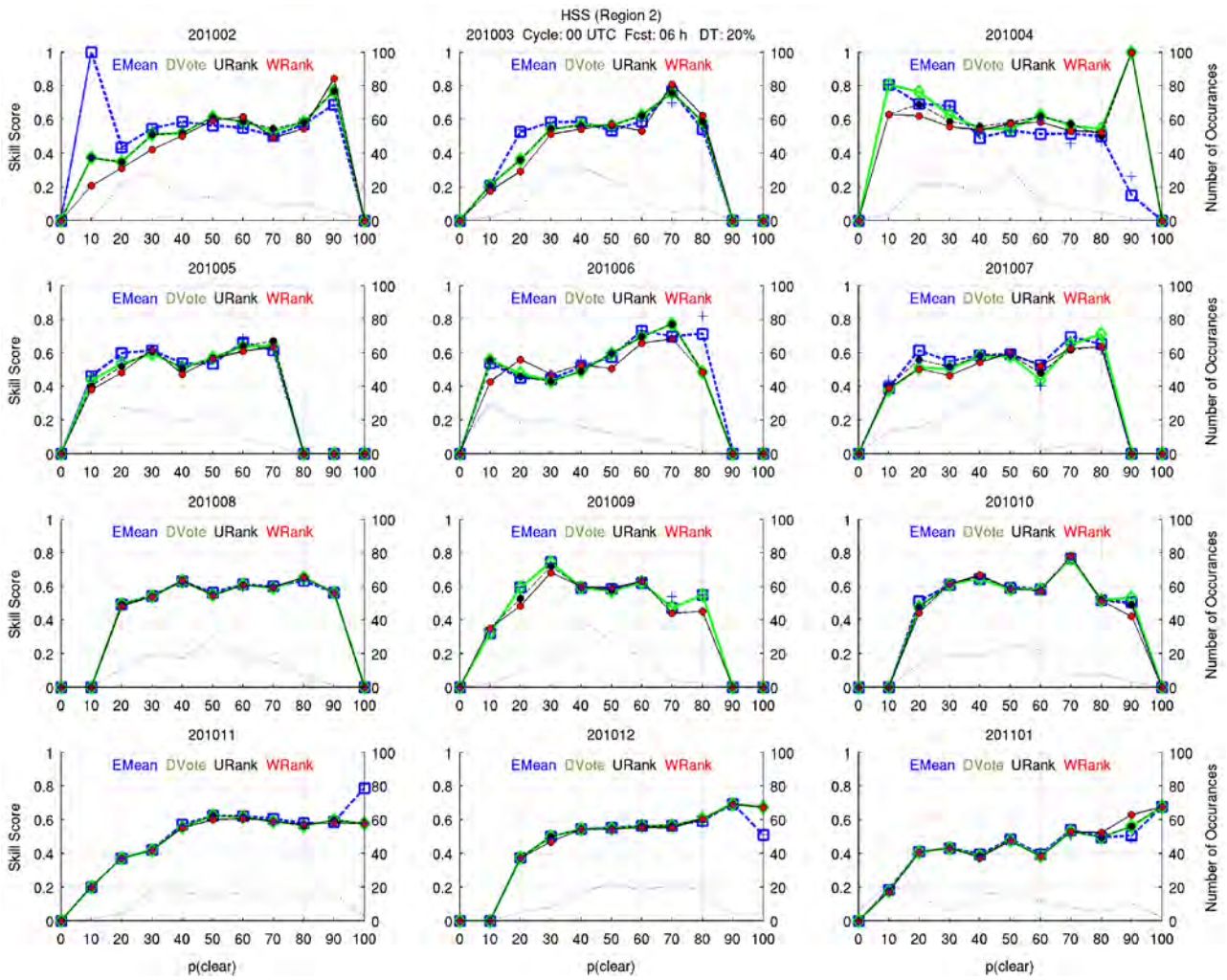


Figure 46. Heidke Skill Score relative to probability of clear conditions (Region 2). Ensemble Heidke Skill Score (left axis) compared to the probability of clear conditions (bottom axis). Thin dashed line indicates the number of grid points (right axis) evaluated for each *p(clear)* threshold (bottom axis).

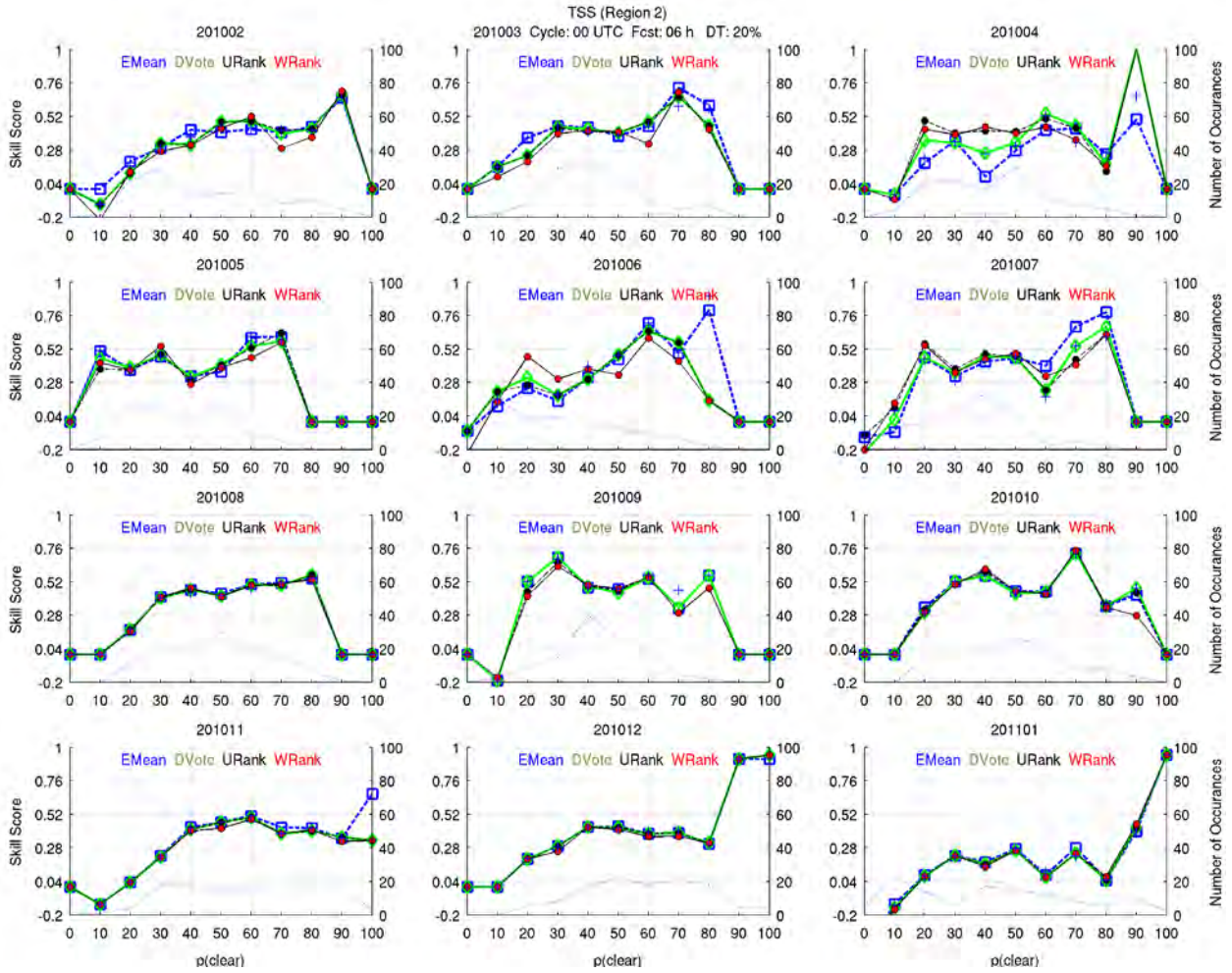


Figure 47. True Skill Score relative to probability of clear conditions (Region 2). Ensemble True Skill Score (left axis) compared to the probability of clear conditions (bottom axis). Thin dashed line indicates the number of grid points (right axis) evaluated for each *p(clear)* threshold (bottom axis).

3. Region 3 (Persistently Cloudy)

a. ROC Diagram

Region 3 has significantly fewer hits, misses, and false alarms than correct rejections. The ensemble probability forecasts have very similar ROC area scores and perform better than the ensemble mean forecast. There is a clustering of skill by the ensemble probability forecasts and separation from the mean from February through July. As with Regions 1 and 2, ensemble probability forecasts generally demonstrate more skill than the mean from February through July. In Region 1, the number of clear events increases during this period in response to the northern deflection of the ITCZ (Figure 28). From August through January, the hit rate decreases steadily as cloud cover increases in response to the southward return of the ITCZ. During this period, however, the ensemble behaves deterministically (August-January).

b. Heidke and True Skill Scores

The distinctions between the ensemble mean and probability forecasts in this region are noticeably different than what we saw in Regions 1 and 2. Here the cloud cover is predominately convective in nature. In this porous cloud region, correct rejections can dominate the forecast outcomes, but the tendency for probability forecasts to predict clear conditions still produces more hits and fewer misses than the ensemble mean. These factors have significant implications on the HSS and TSS measurements of forecast performance.

The ensemble probability, mean, and control forecasts demonstrate significant skill from 10%–90% probability of clear conditions per their HSS results (Figure 49). Differences in skill between the democratic voting and uniform ranks methods are rare. The ensemble mean forecast consistently provides predictions with greater skill at 10% probability of clear than the probability and control forecasts for the first six months – except for the month of June when the control performs best.

Although the ensemble mean's HSS is high at 10% probability of clear, the TSS is less impressive (Figure 50). The large number of correct rejections and the absence of hits, false alarms, and misses reduce the skill significantly. We even see zero skill demonstrated by the ensemble mean in February and May and negative skill in June.

Comparing the HSS and TSS we discover that the skill in this instance has more to do with the lack of clear collection opportunities than with forecast performance.

The differences in HSS between ensemble mean and probability forecasts and mean for 20%–50% probability of clear appear random, but the democratic voting and uniform ranks methods appear to consistently perform better than the weighted ranks method. This is due in large part to the number of false alarms produced by the weighted ranks method brought on by its clear tendency. However, the extreme number of cloudy events that occur in this region at probabilities less than 50% makes the HSS a less than optimal measure of skill.

The TSS, in this environment, measures the ability of the forecasts to produce hits and reduce misses. Therefore, we see that probability forecasts are superior to the mean and control forecasts when probabilities of clear are lower than 50%. The largest disparity between the ensemble forecasts takes place at 20% probability of clear. In this environment where clear conditions are extremely rare, the weighted ranks method consistently performs best. We also note that this extreme difference in skill occurs primarily at peaks in the sample size. This suggests that skill is tied to sampling error. Increasing the number of cloudy events (correct rejections) increases the significance of hits and misses in the TSS computation. Therefore, a larger data set may yield more distinctive results.

As cloud cover frequency becomes more balanced across the region from May through July, the ensemble mean performance improves. The convective nature of the cloud cover does not favor clear or cloudy tendencies. The ensemble mean tends to produce more misses than probability forecasts, which suggest a cloudy bias. The ensemble probability forecasts tend to produce more false alarms than the mean forecast, which confirms its clear bias. This clear bias, however, produces more hits at higher probabilities of clear, which can be useful if timed with the movements of the ITCZ.

The convective nature of Region 3 predominantly produces porous cumuliform cloud cover rather than widespread cloud cover. The TSS is preferred over the HSS at lower probability of clear because correct rejections are so numerous. Although the tendency for ensemble probability forecasts to forecast clear produces false

alarms at a higher rate than the ensemble mean and control forecasts, they are preferred because they also tend to produce fewer misses, and fewer false alarms are produced compared to the number of correct rejections. At higher probabilities of clear, probability forecasts produce more hits than the mean and control forecasts. However, the relatively large number of misses produced by the mean and control suggest that they may have a moist bias.

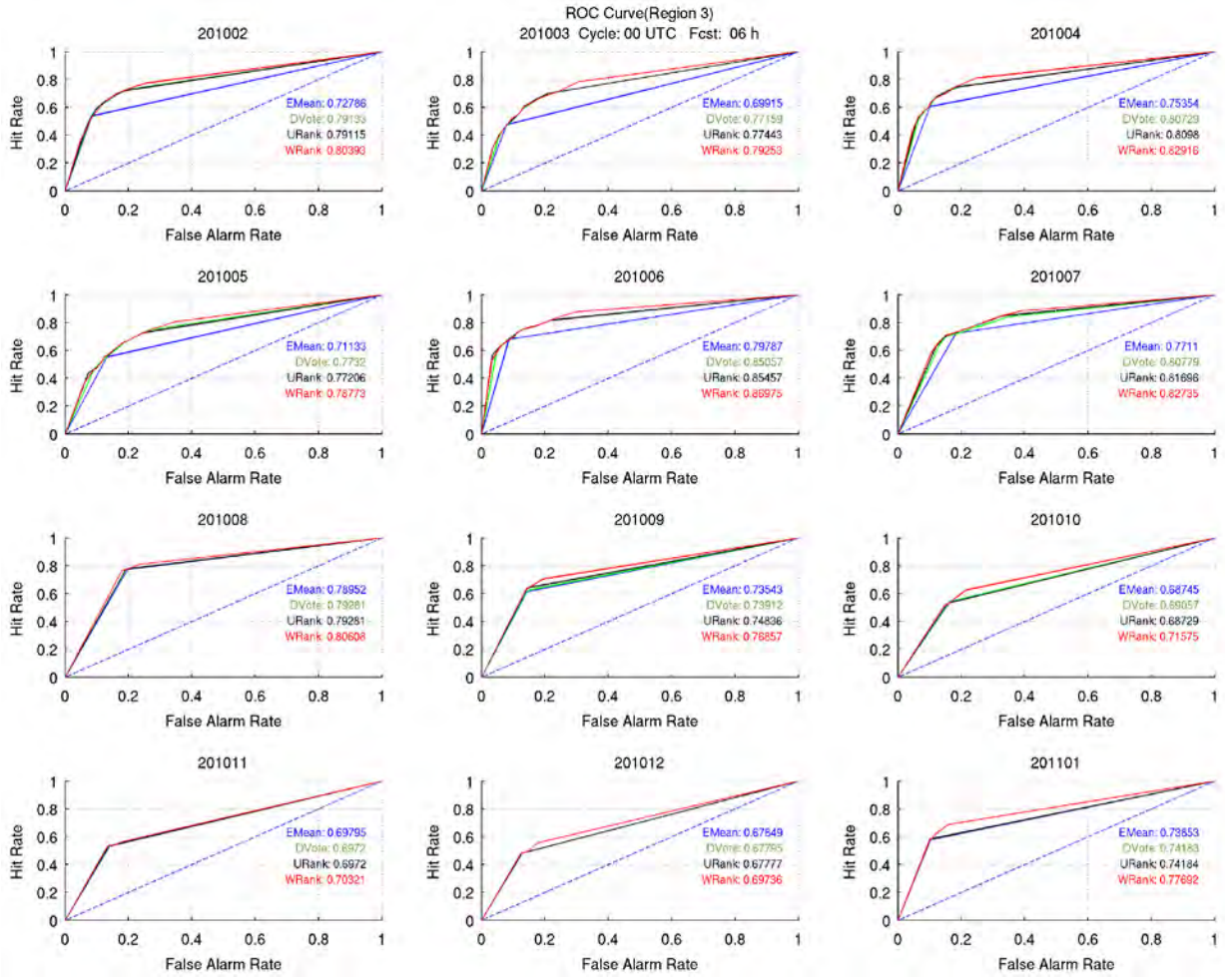


Figure 48. ROC diagrams for Region 3. ROC diagram for 6-h forecast and 0000 UTC cycle. The hit rate (left axis) and false alarm rate (bottom axis) are plotted for the ensemble mean, democratic voting, uniform ranks, and weighted ranks method, and the their ROC areas are included for numerical comparisons.

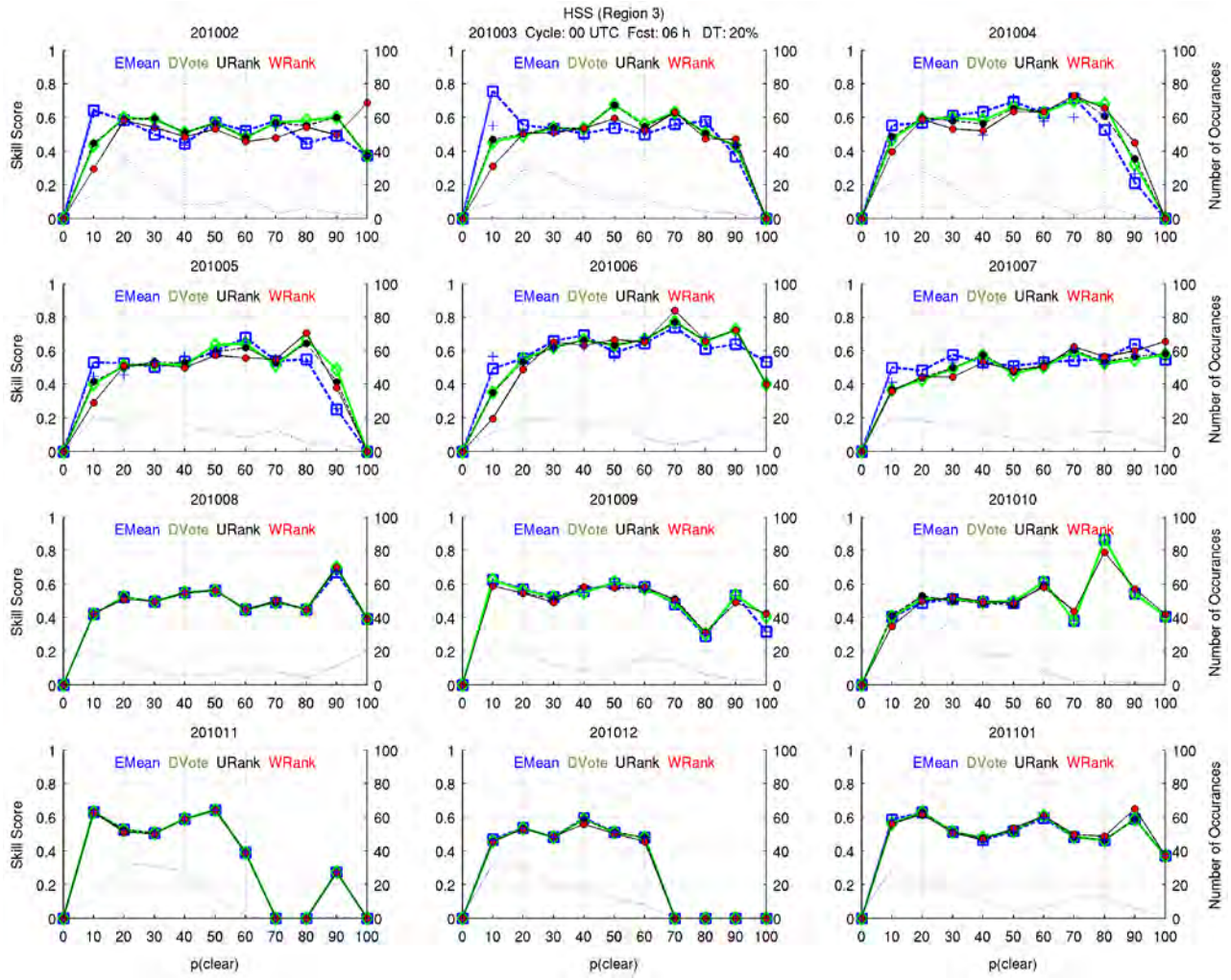


Figure 49. Heidke Skill Score relative to probability of clear conditions (Region 3). Ensemble Heidke Skill Score (left axis) compared to the probability of clear conditions (bottom axis). Thin dashed line indicates the number of grid points (right axis) evaluated for each *p(clear)* threshold (bottom axis).

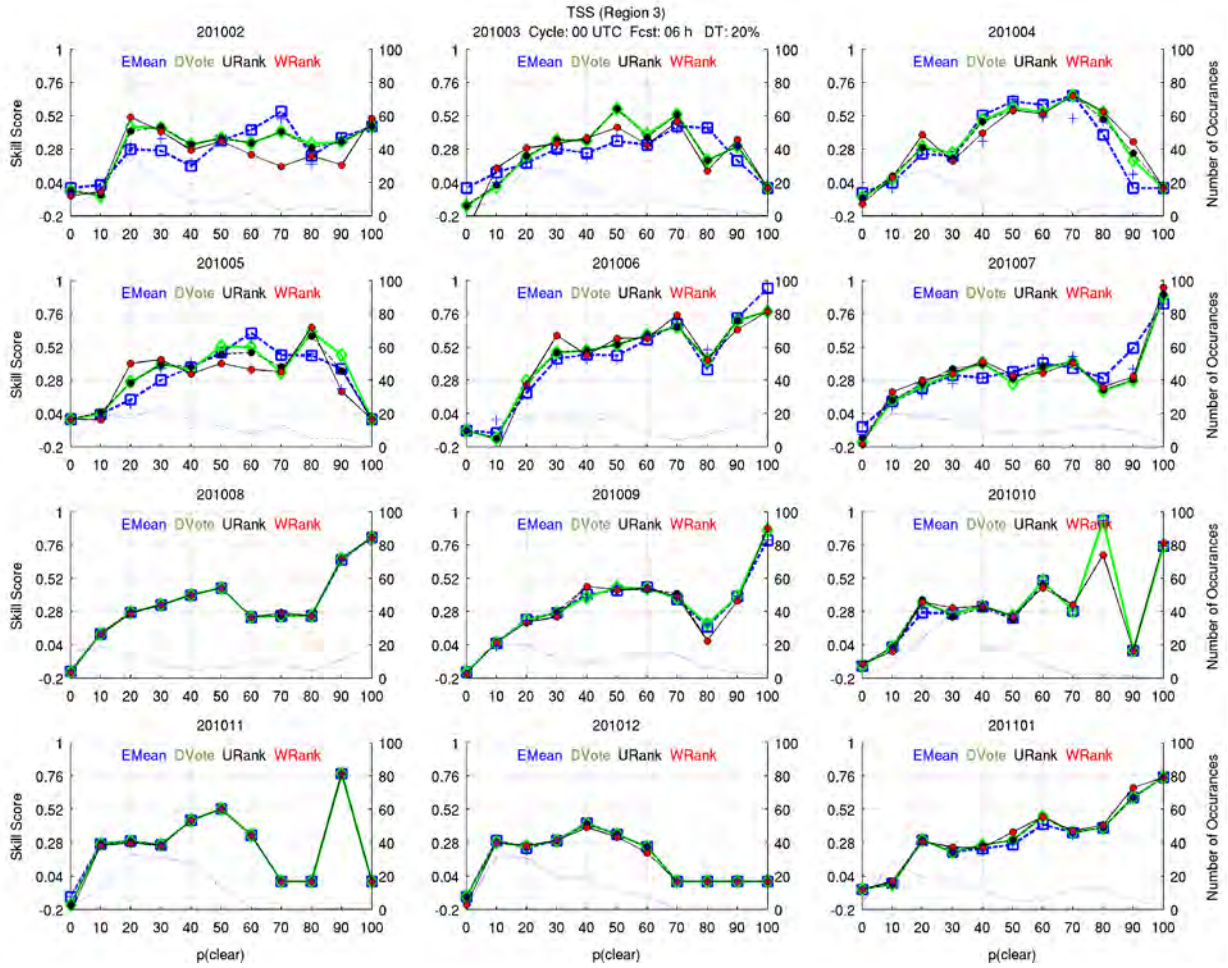


Figure 50. True Skill Score relative to probability of clear conditions (Region 3). Ensemble True Skill Score (left axis) compared to the probability of clear conditions (bottom axis). Thin dashed line indicates the number of grid points (right axis) evaluated for each *p(clear)* threshold (bottom axis).

VI. ENSEMBLE BIAS

Our evaluations of regional skill lead us to three primary questions. 1) Does the ensemble mean forecast truly have a moist bias? 2) How do biases in the ensemble mean forecast compare to the dry bias of probability forecasts? 3) How significant is the ensemble spread? We address these questions in sequence.

First, we will show that the ensemble mean does in fact have biases (mean error), and the magnitude of these biases vary with season and region. Until now, we have made inferences about biases in the ensemble mean forecast. These conjectures are made solely on tendencies seen in forecast outcomes. The mean forecast consistently produces correct rejections and misses at a higher rate than probability forecasts. However, we showed that the ensemble probability forecasts exhibit a clear bias. Thus, the moist bias attributed to the ensemble mean could be a reflection of the ensemble probability forecast statistics. Therefore, we examine the ensemble mean forecast performance independent of other forecasts metrics.

Next, we evaluate the hit, correct rejection, and odds ratio of the ensemble forecast methods. We look at the hit ratio to gauge forecast dependability when clear conditions are predicted. The correct rejection ratio speaks to the dependability of cloudy forecasts. We gather from this measure the likelihood of the forecast to miss a clear event when cloudy conditions are forecasted. The odds ratio used in our skill evaluations is fundamentally the ratio between the hit ratio and the inverse of the correct rejection ratio. Therefore, it articulates the odds of the forecast collecting a clear image versus the odds of missing one. The odds ratio skill score is used instead of the odds ratio to obtain a skill score between 0 and 1.

A. ENSEMBLE MEAN BIAS

Figure 51 shows the monthly biases for the ensemble mean forecast. Biases are calculated according to forecast hour. The differences between the ensemble mean forecast and the analysis corresponding to the forecast verification time are summed and

divided by the total number of forecasts. The ensemble mean forecast clearly indicates a moist bias, but the magnitude of the bias depends on region, forecast hour, and season.

In Region 1, the biases in cloud cover per forecast hour do not appear to result from day-night time analysis development. The 0000 UTC cycle initializes at night (see time zone conversion at Table 9). Therefore, we expect a negative bias in response to the lack of optical capabilities in the nighttime initialization as opposed to the availability of visible imagery in the production of the verifying analysis, but the positive bias suggests the moist tendency in Region 1 is inherent to the ensemble and/or may be an artifact of persistent cloud cover within WWMCA, in the absence of newer satellite imagery, during forecast initialization.

The ensemble mean forecast maintains a moist bias for all forecast hours from February through August. However, we see a decrease in bias at the 12-h forecast lead-time (1600 local). This could indicate afternoon convection, which enhances cloud cover associated with transient systems. Increased cloud cover reduces the disparity between the analysis and the ensemble mean. The largest biases are associated with the 18-h (2200 local) and 24-h (0400 local) forecast lead-times when convection is less active. The inability to employ optical sensors at these times perhaps causes an underestimation of the cloud cover in the analysis. This subsequently magnifies the cloudy bias.

From July through August when most of the observed cloud within the region is located in the southeast quadrant, the ensemble mean continues to indicate a moist bias. However, the hourly variations of the bias converge to ~10% cloud coverage. The 6-h and 12-h forecast biases increase over previous months, but the 18-h and 24-h forecast biases show no significant change. In the midst of persistent cloud cover, the ensemble mean appears to over-predict the amount of moisture in the forecast.

After September when the region becomes most often clear, the cloudy bias subsides. As the number of clouds in the forecast initialization becomes few, the ensemble also forecasts fewer clouds. The 6-h forecast indicates a steady decline in the

moist bias and shows no bias by December. The 12-h forecast shifts to a dry bias until December. The bias of the 18-h and 24-h forecasts decrease through September, but shows an upward trend after October.

In Region 2, from February through August, the ensemble mean demonstrates small cloud cover tendencies that shift slowly between a dry and moist bias. The ensemble is initialized during the day (0800 local), but the biases become increasingly negative with forecast hour. This suggests that the visible imagery used to produce the initial cloud fields do not impact the verification of the forecast, which uses analyses that are produced at night.

Although each forecast hour demonstrates a tendency towards clear conditions in February, we see an upward trend in the bias with each passing month. This progression towards a moist bias coincides with the spring time increase of cloud-cover within the region. By May, the forecast demonstrates a moist bias for all forecast hours. However, the biases of the latter forecast hours remain drier than the 6-h forecast. The fact that the nighttime verification of the 12-h (2000 local) and 18-h (0200 local) forecasts produces a drier bias than the 6-h forecast again suggests that the difference in the day-night WWMCA production has little effect on biases seen in forecast verification.

The biases of the ensemble mean follow the transitions of the South East China monsoon. As monsoonal weather arrives, the ensemble develops a moist bias, which persists through September for all forecast hours. Unlike other forecast hours that maintain a relatively low moist bias throughout the period, the moisture content of the 24-h forecast continues to rise until it peaks in August. After August, monsoonal cloud cover decreases, and the moist bias begins to subside. By November, the ensemble mean forecast biases reach pre-monsoonal values.

In Region 3, the magnitude of the biases between forecast hours show two intelligible patterns. The 6-h forecast is consistently drier than the 12-h forecast, and the 18-h forecast is consistently drier than the 24-h forecast. These patterns persist, independent of seasonal cloud cover.

In February through June, the ITCZ is displaced to the north of the region. During this time most ensemble forecasts indicate a moist bias. However, we see that the 18 h (1400 local) forecast is unbiased from April through July. It is at this time that afternoon convection increases the cloud cover of the analysis and reduces the moist bias within the forecast. The 12-h (0800 local) and 24-h (2000 local) forecasts do not occur at the peak convection times, thus they have strong cloudy biases.

As the ITCZ shifts south, forecasting clear conditions become increasingly difficult. As cloud cover within the analysis increases, the moist bias decreases for all forecast hours except for the 18-h forecast through the month of August. In August, the 18-h forecast develops a moist bias. August also begins the unintelligible behavior of the forecasts. The 6-h forecast maintains a relatively low moist bias. The 12-h forecast develops a relatively high moist bias. The 18-h forecast develops a monthly fluctuation between moist and unbiased predictions. The 24-h forecast biases resemble the biases seen during the northern deflection of the ITCZ.

We conclude that the biases are not regionally specific but specific to the type of cloud cover within each region. Indications suggest that a moist bias exists even when cloud cover is at a minimum as in the dry season of Region 1. With widespread, monsoonal cloud cover, the ensemble moist bias increases as with the monsoon seasons of Regions 1 and 2. The moist bias of the ensemble can be masked in the presence of convection. As convection increases in Regions 1 and 3 the moist bias decreases. The magnitude of the ensemble mean bias depends primarily on the prevailing cloud cover type.

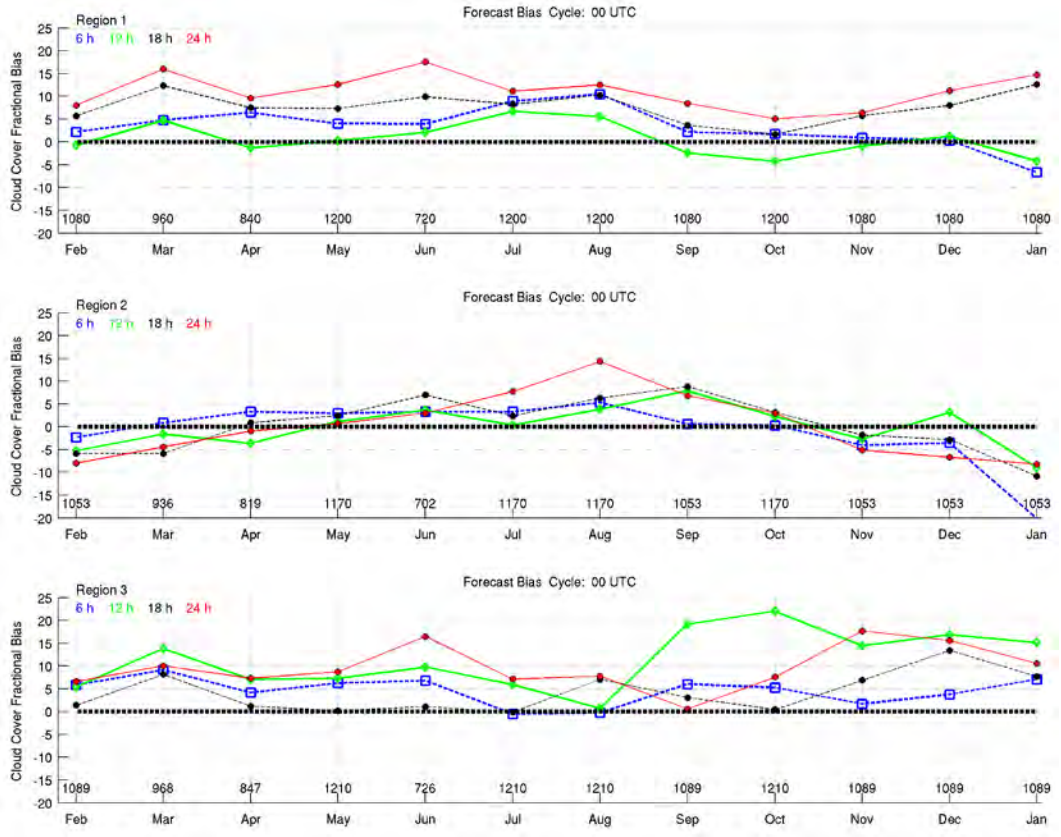


Figure 51. Ensemble Mean bias charts for Regions 1, 2, and 3. The ensemble mean forecast is compared hourly to the analysis (WWMCA) and the monthly bias is plotted.

B. ENSEMBLE SPREAD

Figure 52 shows the frequency in which clear conditions are forecasted according to probability thresholds. It is apparent from the u-shape of the distributions that the ensemble rarely forecasts probabilities other than 100% clear and 100% cloudy. The distribution can occur for several reasons. We have not isolated nor verified the most common causes, but have identified five possible reasons for the tendency for the undersized spread of the ensemble. Four are consequences of the initialization algorithm, and the fifth results from the advection scheme.

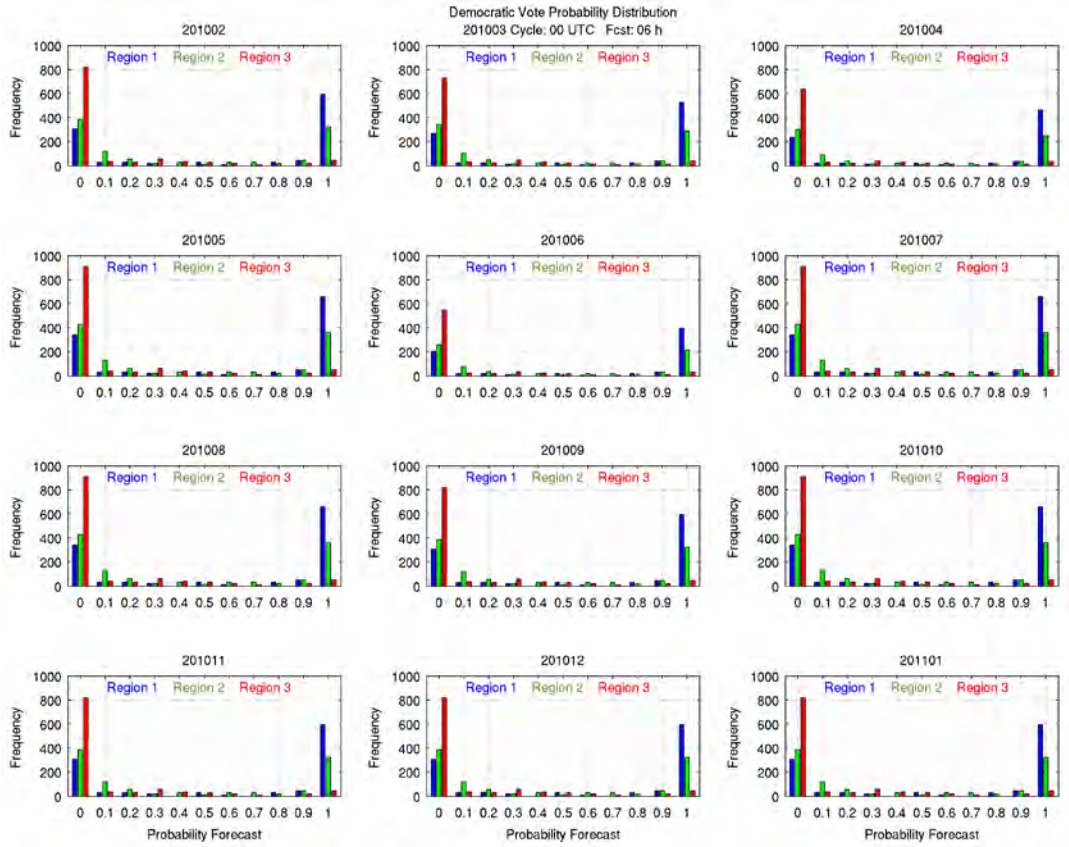


Figure 52. Probability Forecast Distribution. Region 1 (blue), Region 2 (green), and Region 3 (red) histograms of the frequency in which the ensemble (using the democratic voting method) predicts clear conditions at a given probability.

Figure 53 illustrates the first three possible interactions between the analysis and the ensemble during model initialization. The red line represents the initial spread of the ensemble. The minimum and maximum value of the ensemble-member analyses (GEFS) are marked by Min and Max, respectively. The labels above the ensemble spread represent WWMCA values and where they could possibly lie in relation to the ensemble spread. The location of WWMCA values relative to the minimum and maximum ensemble-analyses values can redefine the size of the ensemble spread.

The ensemble spread of the 00-h forecast of ADVCLD is obtained by choosing the driest value between WWMCA and GFS moisture. This becomes problematic in

converting ADVCLD into an ensemble. We return to Figure 53 to clarify the three initialization possibilities. 1) When WWMCA values fall below the minimum ensemble member, the ensemble spread is reduced to zero because all ensemble members adopt the WWMCA value. 2) When WWMCA values fall within the ensemble spread, the spread is reduced, as indicated by the dashed line, and all ensemble members above the WWMCA value adopt the new maximum cloud cover as redefined by WWMCA. 3) When WWMCA values fall above all ensemble members, the spread is unchanged, and the lack of diversity among ensemble members is singularly a property of the initial perturbations of GEFS.

Figure 53. WWMCA Modification of Ensemble Spread. The ensemble spread (red line) is modified when the WWMCA cloud fraction fall below the value of the maximum ensemble-member analyses (Max).

These possibilities are not equally plausible. Because there are a significant number of cases where the ensemble predicts 100% clear, reason 1 could reasonably occur if WWMCA has a dry tendency or the ensemble has a moist bias. The inclination of WWMCA to tend towards extreme cloud fractions makes reason 2 unlikely. Reason 2 can only occur on rare occasions when WWMCA values are other than 0% and 100%. Reason 3 becomes a possibility if WWMCA has a moist tendency. These three possibilities limit the ability of GACE to adequately represent the uncertainty in cloud cover forecast.

Figure 54 illustrates reason 4, which could possibly cause the lack of diversity in the ensemble forecasts. It too is a function of model initialization. When initialization takes place a night (c) and a low WWMCA level does not exist, the cloud fraction is derived from GEFS (reason 3 applies). If there is a low level WWMCA level at night, the driest value between the GEFS and WWMCA is used (reason 1, 2, or 3 applies). When a WWMCA level is identifiable but obscured (b), the standard method for cloud is also used (reason 1, 2, or 3 applies). Reason 4 occurs when the WWMCA level is completely visible as with high clouds (a.). Comparisons are omitted and the WWMCA value is used exclusively. The ensemble spread subsequently goes to zero and tends towards cloudy conditions.

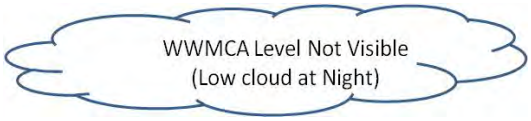
a.		$Cloud = WWMCA\ Lvl$
b.		$WWMCA\ Lvl < MaxLayer$ $WWMCA\ Lvl > 100 - MaxLayer$
		$Cloud = \min(MaxLayer, NWP)$
c.		$Lvl = 0$
		$Cloud = NWP$
		$Lvl \neq 0$
		$Cloud = \min(MaxLayer, NWP)$

Figure 54. Cloud cover depiction of NWP initialization. When the WWMCA Lvl (level) is completely visible (a), the WWMCA value is used to defined the cloud cover amount. When the WWMCA Lvl is obscured (b), the driest value between the ADVCLD Layer (MaxLayer) and the NWP model is used. When the layer is not visible (c), the NWP value is used.

The fifth possible reason is that the curve used to parameterize of cloud cover during the CPS conversion process (Figure 6) can reduce the ensemble spread. Some ranges of the CPS-to-Cloud curve are flatter than others. The variable slopes of the curve can reduce the uniqueness of ensemble members at some cloud fractions and increase the diversity at other cloud fractions.

The skill of the ensemble probability forecasts depends on diversity in the forecasts. The agreement between the ensemble-member forecasts often causes the probability forecasts to resemble the mean. When all ensemble members predict 0% cloud cover, the probability of clear is 100%, and the mean forecasts clear. When all members predict 100%, the probability of clear is 0%, and the ensemble mean forecasts cloudy. Significant differences between the ensemble mean and probability forecasts can only occur if diversity is found among the ensemble members, and the diversity in the ensemble-member forecasts cross the 30% cloud fraction threshold.

The lack of diversity in the ensemble member predictions can be a function of the GEFS and WWMCA. The ensemble initial conditions are perturbed to maximize performance at extended lead-times (4-5 days). Therefore, the initial spread may not be sufficient to capture the desired uncertainty at our 24-h lead-time. The WWMCA day-night analysis process does not appear to contribute significantly to the biases of the mean, but the persistence of upper-level cloud cover information can lead to a moist (dry) bias in the initialization and forecast verification. The extent of these limitations has been considered in this research but not fully examined.

C. RATIO EVALUATIONS

1. Hit Ratio

Figure 55 shows the results of our hit-ratio test. The hit ratio is the total number of hits divided by the total number of clear forecasts. The monthly hits and false alarms are tallied for each region and the hit ratio is calculated. The hit ratio communicates the probability or likelihood of collecting a clear image when the clear conditions are

forecasted. High numbers are preferred. To maintain consistency with previous analysis, we evaluate the 0000 UTC cycle, 6-h, and 20% decision threshold forecast.

In Region 1, where opportunities to collect clear imagery are many, the hit ratio is substantially high. The ratio increases from February to May as frontal systems moving through the area become less frequent. During the summer months (May-August), the hit ratio becomes relatively stationary for all forecast methods.

The ensemble mean and control forecasts maintain a higher hit ratio than the probability forecasts. Although the ensemble mean has a moist bias, it provides the most skill in producing hits and avoiding false alarms. The weighted ranks method produces the worst hit ratio in this region. We note, however, that the difference in the hit ratios for all forecast methods is less than 5% from February to May and even smaller during the summer months.

In Region 2, the small number of opportunities to collect cloud-free imagery along with the variability of the cloud cover in the region makes the hit ratio less impressive. The ensemble mean and control forecasts perform better than the probability forecasts. In the case of the weighted ranks method, the difference in the hit ratio can be greater than 10%.

The cloud-cover variability of the region also engenders separations in the hit ratio of forecast methods that are usually indistinguishable in performance. The hit ratio of the ensemble mean forecast maintains relatively stationary from February through July. The hit ratio of the control forecast, generally equivalent to the ensemble mean, demonstrates less skill in March, May, and July. As cloud cover becomes more variable (Feb–Apr), the hit ratio of probability forecasts increases. The enhanced cloud cover associated with the summer monsoon, however, causes the hit ratio of the probability forecasts to decrease.

In Region 3, the distinction between the ensemble mean and other forecasts is very prominent. The prevailing cloud cover in the region makes clear forecasts rare. Therefore, false alarms significantly decrease the hit ratio. In this region, the mean

forecast provides the best opportunity to minimize false alarms. We also see that the moist bias in the control forecast is not as strong as the ensemble mean, so it produces more false alarms. The ensemble forecast produces a significant amount of false alarms with the weighted ranks method being the worst.

Operators who are interested in collecting imagery but are equally concerned about avoiding cloudy imagery should prefer the ensemble mean forecast. For each region, we find that the hit ratio of the control forecast is higher than all other forecasts. Although the differences in Region 1 are small, the ensemble mean produces better results. The variability in cloud cover seen in Region 2, increases the chances that a prediction of clear conditions will have a negative outcome. Therefore, extremely cautious users should use the ensemble mean forecast. In like manner, the persistent cloud cover in Region 3 makes for an environment suitable for the ensemble mean to produce superior performance in avoiding cloudy collections.

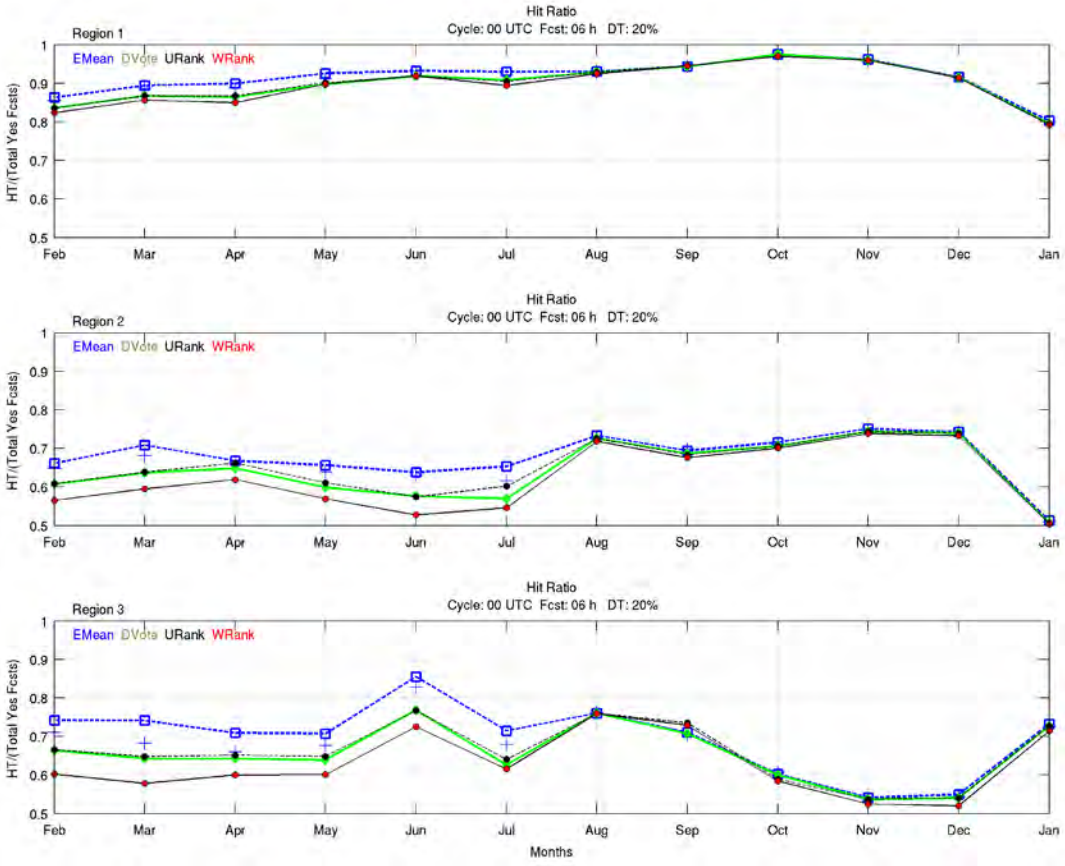


Figure 55. Hit ratio (hits divided by the number of "yes" forecasts) calculated for Regions 1, 2, and 3. Ratio plotted monthly for ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) methods.

2. Correct Rejection Ratio

Figure 56 shows the correct rejection ratio results. The correct rejection ratio is the total number of correct rejections divided by the total number of cloudy predictions. The monthly correct rejections and misses are tallied for each region and the correct rejection ratio is calculated. The correct rejection ratio communicates the probability or likelihood of correctly aborting a cloudy image when cloudy conditions are forecasted. High numbers are preferred. We can also gather from this measure the likelihood of the forecast to miss a clear collection. To maintain consistency with previous analysis we

evaluate the 0000 UTC cycle, 6-h, and 20% decision threshold forecast. The differences between the regions are not significant; therefore, we discuss them together.

The correct rejection ratio varies little from month to month. Outside a 10% decrease in Region 1 from February to March, the correct-rejection ratios remain relatively stationary. The correct-rejection ratios of the ensemble forecast methods rarely differ more than 10%. The ratios are highest in Region 2 and 3 because opportunities to collect clear imagery are rare.

The weighted ranks method produces the highest correct rejection ratio. The tendency for ensemble probability forecasts to predict clear conditions reduce the number of missed collection opportunities as compared to the ensemble mean and control forecasts, which preference cloudy conditions. Although the correct rejection ratio differences between ensemble forecast methods are small, users who are extremely sensitive to missing a collection opportunity should use the weighted ranks method.

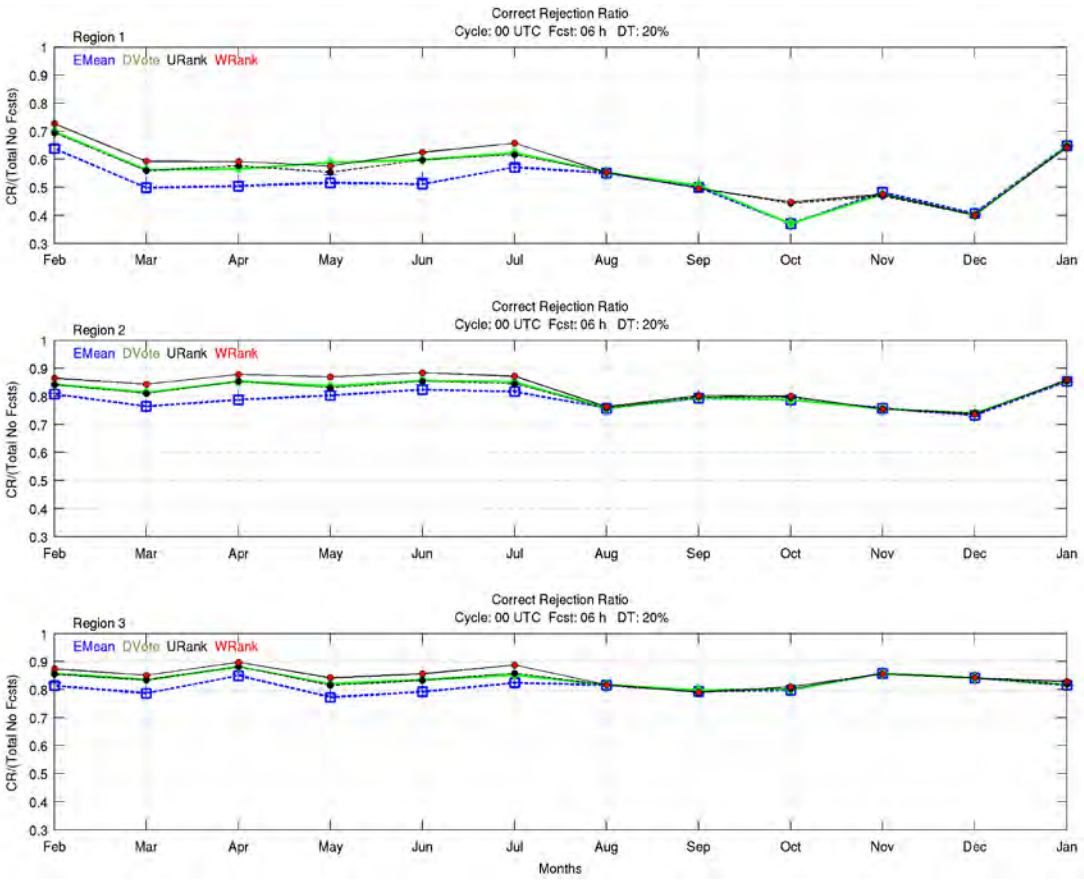


Figure 56. Correct rejection ratio (correct rejections divided by the number of "no" forecasts) calculated for Regions 1, 2, and 3. Ratio plotted monthly for ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) methods.

3. Odds Ratio

Figure 57 shows the odds ratio skill score ORSS results. The ORSS combines the hit and correct rejection ratios to produce a ratio that describes the probability or likelihood of collecting a clear image versus missing a collection opportunity. High numbers are preferred. To maintain consistency with previous analysis we evaluate the 0000 UTC cycle, 6-h, and 20% decision threshold forecast.

In Region 1, the performance of the forecasts is seasonal. From February through March when frontal system move through the region, the ensemble weighted-ranks

method should be preferred. From April through May when frontal system move through less frequently, the ensemble mean forecast should be preferred. During June, the transition between the predominant clear condition and the South Asian Monsoon, the weighted ranks method should be chosen. The ensemble mean forecast should be preferred when the monsoonal clouds over fully impacts the region.

In Region 2, there are also seasonal trends in forecast performance. From February through March, slight differences exist between forecast methods. In April when the monsoon season begins, the probability forecasts perform significantly better than the ensemble mean and control forecasts. The weighted ranks method produces the best odds of collecting a clear image versus missing a clear image from the beginning of the monsoon season (April) to June. As the max rainfall moves over the region in July, the ensemble mean forecast performs slightly better. Prior to the onset of the monsoon season, the weighted ranks method should be the preferred forecast. After the onset of the monsoon season, however, the ensemble mean should be the preferred forecast method.

In Region 3, the likelihood of collecting a clear image versus missing an image varies from month to month. The ORSS difference between the ensemble forecast methods are largest in February and March when the ITCZ is located in the southern portion of the region. During this period, the ensemble mean forecast performs best, and the weighted ranks method produces the worst results. The control forecast demonstrates less skill than the ensemble mean and is also surpassed in skill by probability forecasts.

The sub performance of the ensemble control forecast continues into April and May when the ITCZ reaches is southernmost deflection and begins to transition back into the region. During this period, all forecasts perform nearly equally except the ensemble control forecast, which performs worse. However, the ORSS is relatively high at ~80% within this extremely convective environment.

Convection decreases across the region as the ITCZ moves to the north in June, and the ensemble mean perform best. The ensemble control does not perform as well as

the mean but is more skillful than the probability forecasts. During this period when cloud-cover becomes increasingly porous, the tendency for the control forecast to predict cloudy conditions produces better results. By July, the ITCZ shifts to the northern border of the region, which results in a slight clearing of the region. At this time, the weighted ranks method shows improved skill.

Due to similarities in the ORSS of the ensemble forecasts, this score cannot lead to a conclusive assessment of which forecast method is most useful. However, we can make some inferences about the skill of the forecast methods. In Regions 1 and 2, we find the clear bias of the weighted ranks method to be most beneficial in regimes where the dominate cloud cover features are transient. Examining Regions 1 and 2 also revealed that the ensemble mean performs best where cloud cover persists, as in monsoonal weather. We discover in Region 3, that the ensemble mean also performs best in environments that are characterized by extreme convection.

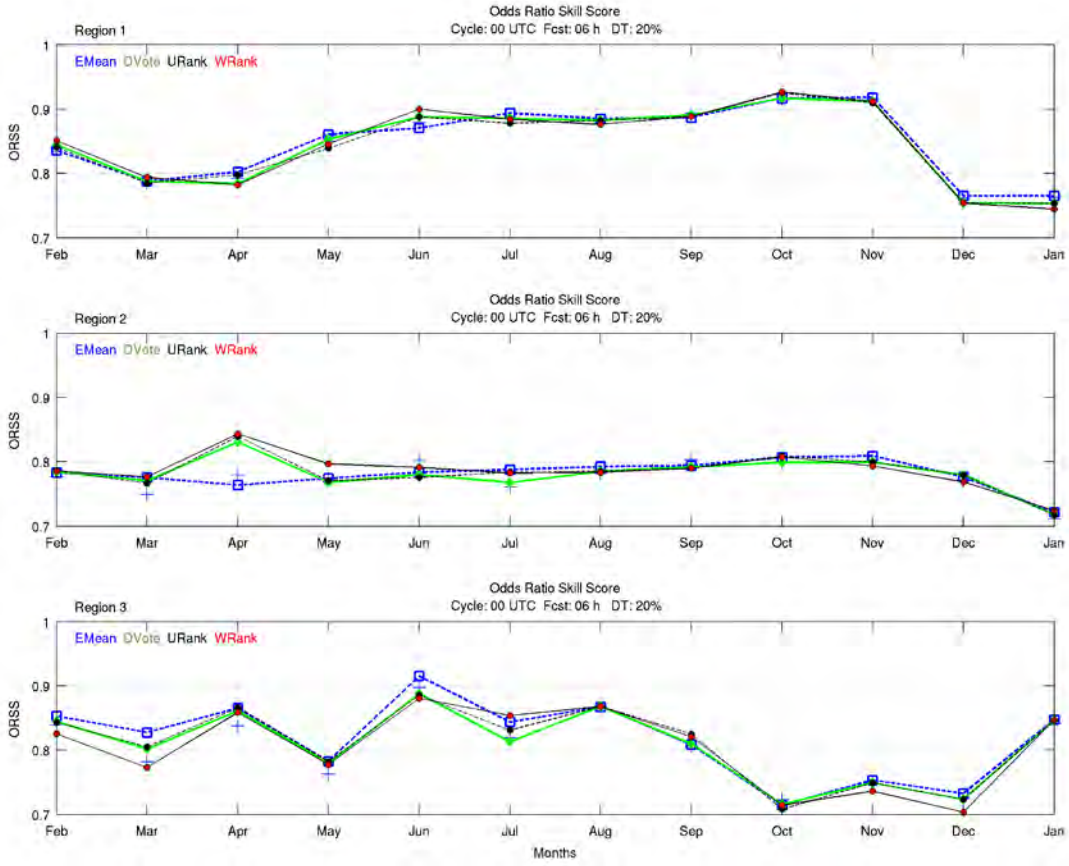


Figure 57. Odds Ratio Skill Score (odds of correct clear forecast vs. incorrect cloudy forecast) calculated for Regions 1, 2, and 3. Ratio plotted monthly for ensemble mean (square), control (+), democratic voting (diamond), uniform ranks (black dot), and weighted ranks (red dot) methods..

THIS PAGE INTENTIONALLY LEFT BLANK

VII. ENSEMBLE UTILITY VALUE

We have shown that the ensemble forecasts demonstrate skill in predicting clear cloud conditions, but skill does not necessarily communicate how useful the forecast information is to the application or operation it supports. Decision makers who discover that an excellent forecast is of little consequence to their operation can be empowered to make decision without heavily weighing the forecast. In the opposite case, the decision maker is ill-advised to make decisions without consulting the forecast. Therefore, it is useful to know how much more effective operations can be with the addition of forecast information.

We chose 20% probability of clear for our decision threshold during our analyses of forecast skill; now, we examine the expected value of the forecasts relative to the entire range of probability decision thresholds (DT). The plots that follow show the expected utility value a user can obtain from following the forecast. We begin with the ensemble forecast outcomes at each probability decision threshold. These probability decision thresholds are mapped directly to user defined sensitivities to cloud cover imagery, the utility of correct rejections *u(CR)*. Next, we calculate the maximum expected utility of climatology. Then, the maximum expected utility is subtracted from the expected utility of the perfect forecast and each ensemble forecast method to compute respective expected utility values.

We examine the expected utility value of the ensemble mean, control and probability forecasts in each region. The 0000 UTC cycle and 06-h forecast are evaluated, as with the skill assessments. However, the charts in this section do not include August through January data. Instead, the figures contain the expected utility value of the risk-neutral and risk-averse users over the first six months of our dataset. We focus our discussion on the risk-neutral and risk-averse cases since these responses to risk are most common. The expected utility value of perfect information (black dashed

line) is plotted to help gauge the difference between the expected utility value of the forecast and the maximum possible expected utility value that can be gained.

A. REGION 1

We begin in Region 1 with the risk-neutral and risk-averse cases (Figure 58). The ensemble mean and control forecasts provide the most value for users who have a low tolerance (DT $\geq$ 70%) for cloud filled imagery. The ensemble probability forecasts provide the most utility to high tolerant users (DT $\leq$ 30%) and indifferent users (40 $\geq$ DT $\leq$ 60). These results resemble our skill evolutions. We discovered that biases in the forecasts are reflected in the HSS and TSS of the ensemble. We also see that these biases have implications on the utility of the forecast relative to user probability decision thresholds.

The moist bias of the ensemble mean is particularly evident in Region 1. The reluctance of the ensemble mean forecast to predict cloud-free conditions at high probability decision thresholds results in fewer cloud filled images. This is utilitarian for users who have a low tolerance for cloud filled imagery but not high tolerant users who are willing to accept more risk to collect a clear image. Therefore, the ensemble mean experiences a decline in forecast utility for high tolerant users.

The dry biases of the ensemble probability forecasts are perhaps more evident than the biases seen with the ensemble mean forecast. The ensemble probability forecasts provide less value than the ensemble mean for users with high probability decision thresholds. The increased number of cloud filled images, which result from the tendency to predict clear conditions, reduces the utility of these forecasts. In addition, the increase in forecast utility becomes non-linear.

The more sensitive the user is to collecting cloud filled imagery the greater the disparity in utility seen between forecast methods. The weighted ranks method has the strongest dry bias; therefore, it provides the least amount of utility. The uniform ranks

method, which adds additional probability to the democratic voting method, provides less utility than the democratic voting method but more than the weighted ranks method.

Risk-neutral users experience significant changes in the utility of each forecast method based on the decision threshold chosen—as determined by the slope of the utility curve. From February through April when cloud cover decreases within the region, users who have a high tolerance for cloudy imagery (DT=10%–20%) are encouraged not to use the forecast. From May through July when cloud-cover increases over the southeastern quadrant of the region, users with a probability decision threshold of 20% begin to see utility in using the forecasts. We see similar trends in the risk-averse case.

Risk-averse users, however, experience smaller changes in expected utility value with differing probability decision thresholds—this is a byproduct of the utility function used to calculate the expected value of the forecasts. All decision thresholds experience an increase in expected utility value over the neutral case in the first three months. In the latter three months, however, the 90% probability decision threshold does not receive a bump in expected utility value. Perhaps the most significant increase in expected utility value of the forecasts happens for users choosing the 10%–30% probability decision threshold. We are careful to note that these increases are not in response to changes in the forecast but in the willingness of the user to accept more risk.

From February through April, changes in the expected utility value of the forecasts are subtle at best. Users who employ the ensemble mean forecast obtain maximum utility value at the 90% probability decision threshold. Those using an ensemble probability forecast gain the most utility value at the 70% probability decision threshold.

From May through July, the increased cloud cover over the southeast quadrant of the region impacts the expected utility value of the forecasts. The difference is most prominent in July. Although the ensemble mean forecast continues to produce the best results for high probability decision thresholds, expected value no longer peaks at the 90% probability decision threshold. The new peak for all forecasts is found at the 50%

probability of clear conditions. In the presence of this increased cloud cover, we also see a drop in expected utility value for all decision thresholds, as compared to the previous months.

In Region 1, neither ensemble forecast method truly separates itself by producing superior expected utility value, but there are notable differences. Users who are risk-neutral must be very sensitive or accurate in choosing their decision thresholds, or utility of correct rejections, because the expected value of the forecast is very responsive to these values. These users are ill-advised to use the forecasts at very low probability decision thresholds. However, the ensemble probability forecasts should be considered at low probabilities, and the ensemble mean forecast should be considered at high probability decision thresholds. Users who are risk-averse gain more utility from the forecasts than risk-neutral users. At the 10%–20% probability decision threshold, risk-averse users gain expected utility not afforded to risk-neutral users.

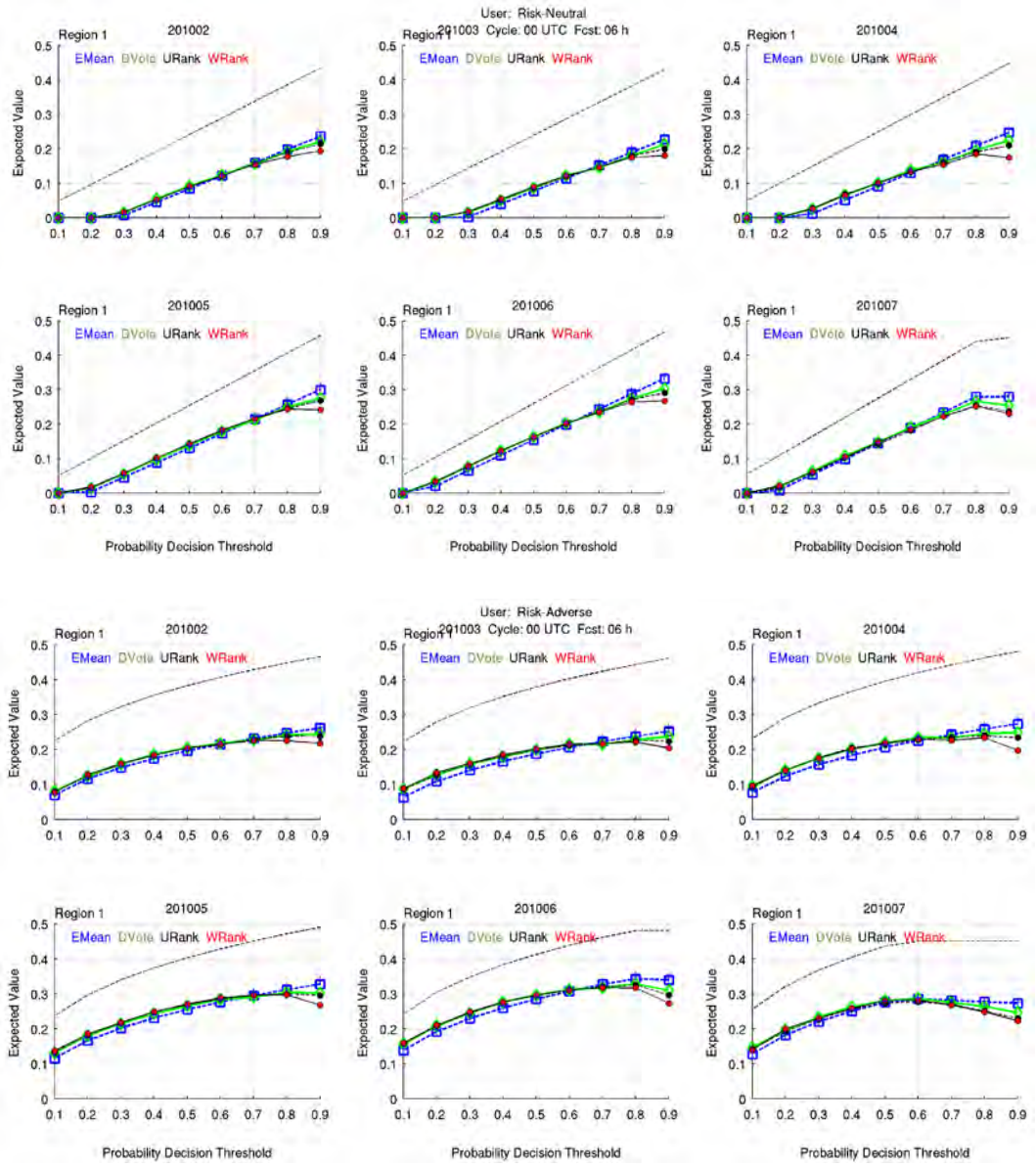


Figure 58. Expected value vs. utility plot (Region1). Error bars represent the standard deviation of each forecast method. Probability decision thresholds (bottom axis) may also represent assigned utility value of cloud-covered imager or a user assigned image priority. Perfect utility value (dashed line) plotted for maximum value attainable.

B. REGION 2

Figure 59 shows the risk-neutral and risk-averse cases for Region 2. The increased cloud cover within the region, as compared to Region 1, leads to a reduction in

the utility of the forecasts across all probability decision thresholds. It also leads to a reduction in the difference between the expected utility values of the forecast methods. Month-to-month changes in cloud cover within the region are reflected in the lack of consistency in determining which ensemble forecast method provides the most expected utility value at high and low probability decision thresholds. Furthermore, biases in the ensemble forecast methods are become more difficult to decipher.

Risk-neutral users experience a shortened range of probability decision thresholds. Positive expected utility value for all forecast methods is limited to probability decision thresholds greater than 30%. In addition, the highly variable conditions make expected utility value differences between the ensemble probability forecast methods indistinguishable.

From March through May, the frequency of clear conditions fluctuates. In April, the ensemble mean forecast provides the most expected value because the number of clear events is reduced from the previous month. In May, the number of cloudy events is reduced and most of the region is clear 50% of the time. During this period of high uncertainty, the ensemble probability forecasts provide more expected value than the ensemble mean and control forecasts for indifferent users.

From June through July, the expected utility value of the ensemble forecasts converges even further with the introduction of monsoonal cloud cover. In June, a substantial amount of the region experiences 50% probability of clear conditions, but the region becomes progressively cloudier. The cloudy conditions in July are reflected in the moist bias of the ensemble mean forecast, which provides the most expected utility value for users in the 30%–50% probability decision threshold range.

Risk-averse users are presented more opportunities to take advantage of the biases of the ensemble forecasts. Those who employ a low probability decision threshold see the greatest difference in expected utility value of the forecasts. At probability decision thresholds from 10%–50%, the ensemble mean forecast provides the most expected utility value except in April. At these decision thresholds, we are not able to distinguish

between the expected utility value of the democratic voting and uniform ranks methods, but we do find discernable differences between these forecasts and the weighted ranks method. The strong tendency for the weighted ranks method to predict clear at lower probabilities produces significant false alarms in this region. This behavior is not evident in April due to the significant peak at 50% probability of clear (climatological).

Those who elect to employ probability decision thresholds greater than 40% will generally not find one ensemble forecast to be any different from the next. The expected utility value of each forecast method has a direct relationship with user sensitivities to cloudy imagery from February through May. However, in June the peak in expected utility value begins to shift toward lower probability decision thresholds. The maximum expected utility value shifts to 60% probability decision threshold as monsoonal cloud cover begins to dominate the region.

In Region 2, cloud cover is variable within and between months. Therefore, the differences in ensemble forecast performs also varies from month-to-month. Risk-neutral users find no expected utility value in following the forecasts at probability decision thresholds below 40%. Although the differences in expected value of the forecast methods are not substantial, they follow seasonal trends. When cloud-cover increases, the ensemble mean forecast provides more utility value. When cloud-cover decreases, the ensemble probability forecasts provide the most utility value. These signals are not strong, but they are indicators of the biases that exist in the ensemble forecast methods.

Risk-averse users naturally find more value in the use of additional information. Decision makers among these users who have a low probability decision threshold should elect to use the ensemble mean forecast. Although the probability forecasts tend to forecast clear conditions at these probabilities, cloudy bias of the ensemble mean forecast provides greater utility in regions where cloud cover is widespread. We also note that ensemble probability forecasts perform better in April, but the ensemble mean forecast performs better on average. Above the 50% probability decision threshold either forecast will yield similar results.

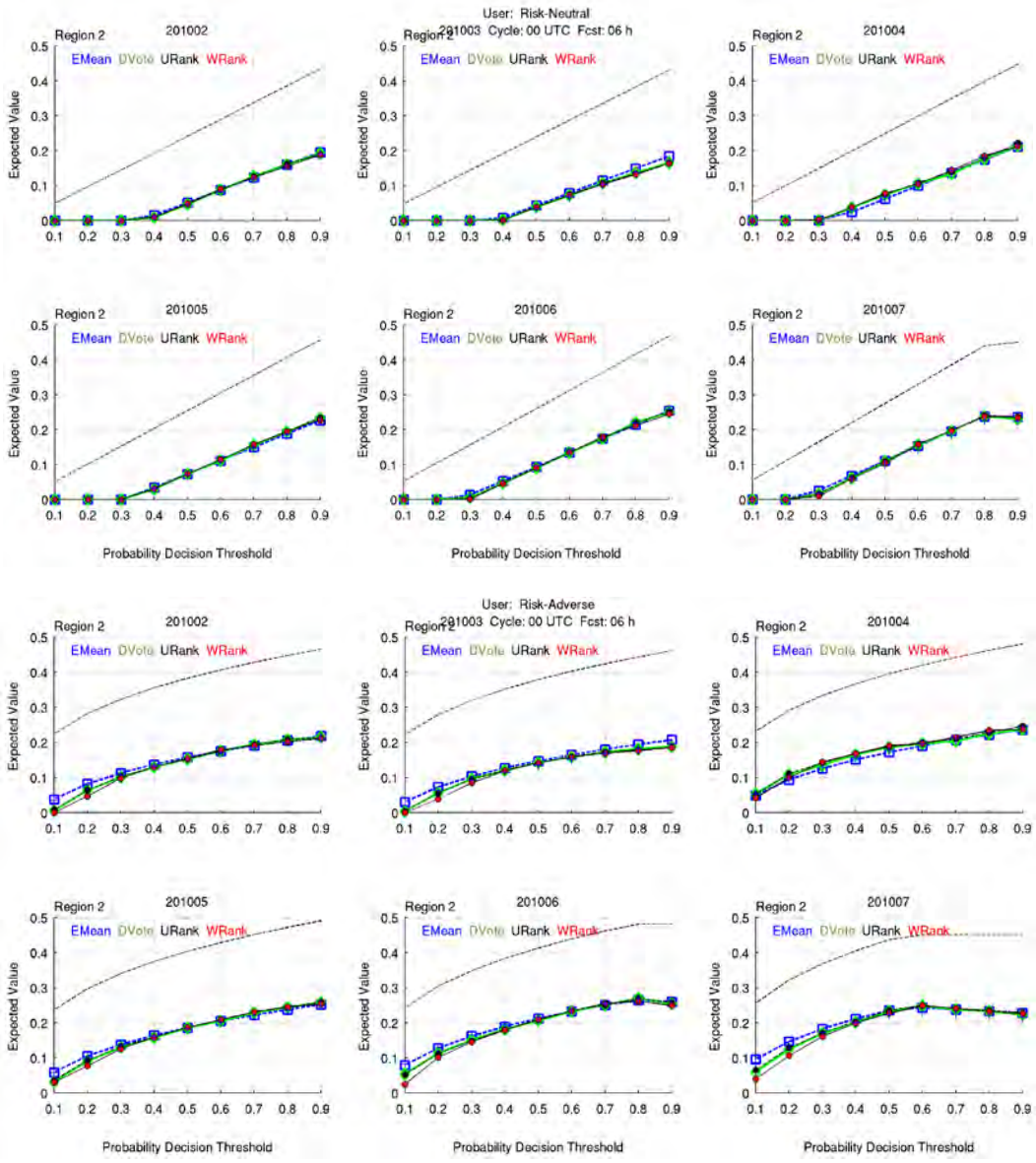


Figure 59. Expected value vs. utility plot (Region2). Error bars represent the standard deviation of each forecast method. Probability decision thresholds (bottom axis) may also represent assigned utility value of cloud-covered imager or a user assigned image priority.

C. REGION 3

Figure 60 shows the risk-neutral and risk-averse cases for Region 3. Although this region is predominantly cloudy, the dynamic nature of the cloud cover makes it a tremendous forecast challenge. The frequency of clear events changes from month-to-

month with the North to South oscillation of the ITCZ, but the overall uncertainty within the region remains the same. All forecast methods have significant challenges distinguishing between cloud and clear events. Never-the-less, both risk-neutral and risk-averse users can obtain utility value by following the forecasts in this region.

Risk-neutral users find little differences between forecast methods. Users who chose extremely low probability decision thresholds receive no value in using the ensemble forecasts. Users employing probability decision thresholds above these extremely low values can receive utility value from any of the ensemble forecast methods. The expected value at most of the probability decision thresholds is the same for each ensemble forecast method. However, the ensemble mean forecast demonstrates more value when differences occur.

The only significant differences between the forecasts occur at probability decision thresholds greater than 70%. Here the probability forecasts, which tends to have a dry bias, produces more false alarms than the moist biased ensemble mean and control forecasts. This appears to be a consistent trend except for the months of April and July. During these months, there are slight changes in the bias of the ensemble mean forecast. In April, the magnitude of the bias temporarily decreases. Therefore, the forecast outcomes become comparable to those seen with the probability forecasts. In July, ensemble mean forecast is unbiased and provides more utility value at lower probability decision thresholds than the ensemble probability forecasts.

Risk-averse users would also be better served to use the ensemble mean forecast. Users who employ the probability forecast will be penalized by the inherent dry bias of the forecast. Because of these forecast biases, the ensemble mean and control forecasts provide equal or better expected utility value than the ensemble probability forecasts for each month.

From February through Jun, the biases are most prevalent at the extreme probability decision thresholds, from 10%–30% and 70%–90%. The weighted ranks method produces the least expected value at these ranges. Although the democratic

voting and uniform ranks methods are subordinate in value to the ensemble mean and control forecasts, the differences are not significant at lower probability decision thresholds. In addition, decision makers will not find significant differences between all forecast methods at probability decision thresholds between the extreme values.

In April and July when the bias of the ensemble mean is reduced, the performance of the mean changes with respect to the probability forecasts. In April, the expected value of the ensemble mean at high probability decision thresholds is not as distinguishable from that of the probability forecasts. The same is true for July when the ensemble mean produces unbiased forecasts. The unbiased forecasts also induce a greater separation in skill at lower probability decision thresholds.

In Region 3, both risk-neutral and risk-averse users should chose to employ the ensemble mean forecast over probability forecasts. The dynamic nature of cloud cover combined with the oscillation of the ITCZ makes forecasting a challenge for all ensemble forecast methods. In the midst of the chaos, however, the ensemble mean produces the most expected utility value when differences between the forecast methods occur.

These differences are found at the extreme probability decision thresholds. Risk-neutral users primarily see an advantage in using the ensemble mean forecast at high probability decision thresholds. Risk-averse users experience an advantage in utilizing the ensemble mean forecast at both high and low decision thresholds. The two exceptions occur in April and July when the bias in the ensemble mean forecast is not as prevalent. During these months, we see a reduction in the expected value at high probability thresholds and an increase at lower probability thresholds, as compared to probability forecast methods.

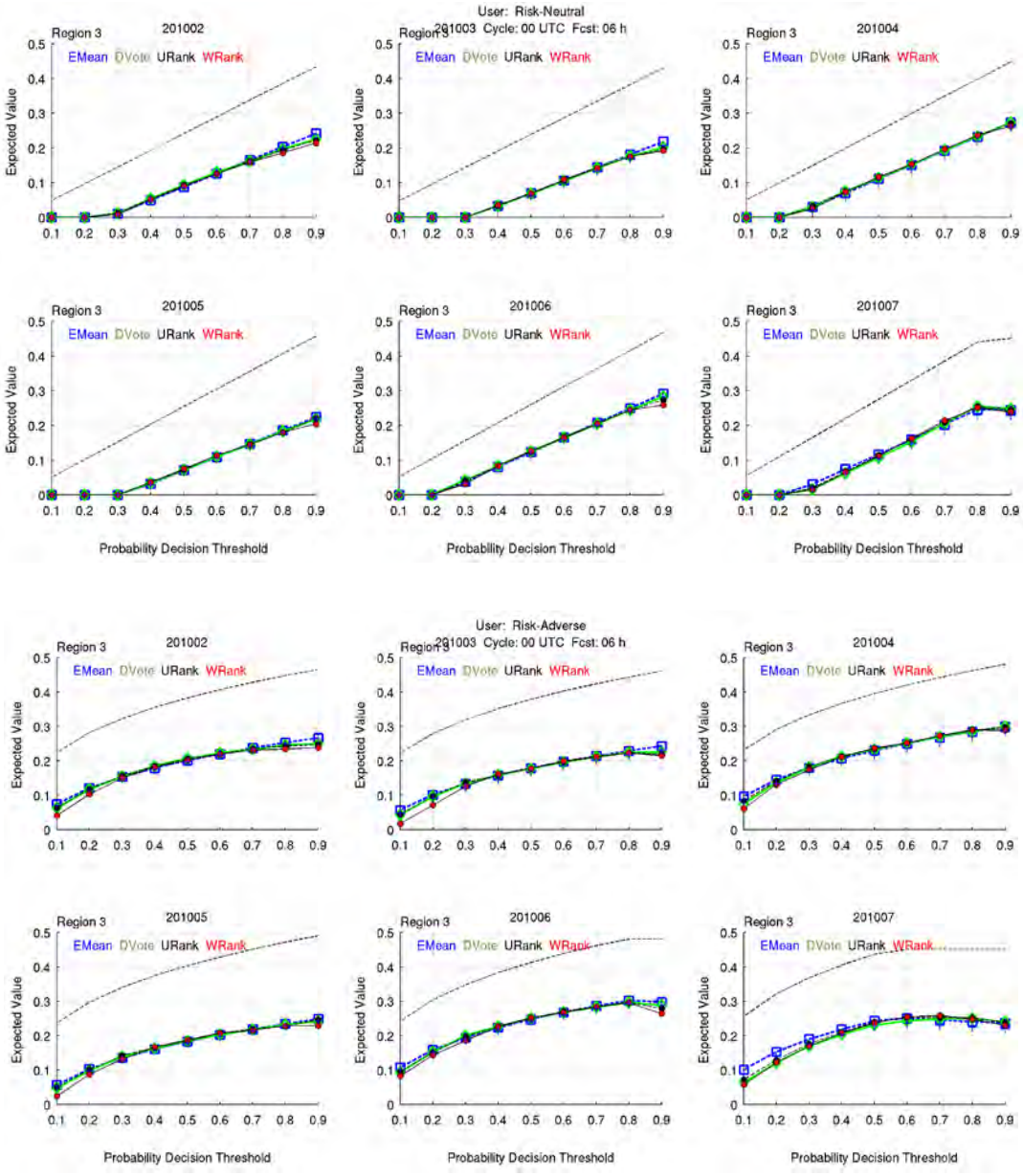


Figure 60. Expected value vs. utility plot (Region3). Error bars represent the standard deviation of each forecast method with respect to the assigned utility value of correct rejection.

D. SUMMARY

Expected utility value evaluations suggest that additional cloud-cover information provided by the five ensemble forecast methods presented can reasonably improve the operational efficiency in capturing cloud-free imagery. The potential enhancement to

collection operations through the use of these ensemble forecast methods varies with user type, cloud cover type, and operation type. Combinations of these criteria can produce an expected utility value as low as 0% (4%) and as high as 34% (32%) with risk-neutral (risk-averse) users.

If we assume that imagery collection operations and operators are not indifferent to cloud filled imagery, we can conclude that a measure of risk exists with all cloud-free collection operations. The level of risk aversion depends on the type of user. In employing the $s^{0.5}$ utility function to define the sensitivities of our risk-averse user, we discover that the ensemble demonstrates the potential to add significant value to cloud-free forecast operations. The ensemble mean and the weighted ranks methods consistently provide the least or most expected utility value relative to the democratic voting, uniform ranks, and control forecasts.

Table 10 summarizes the expected utility value results for the ensemble mean and weighted ranks methods. The expected utility value of the other three methods falls somewhere between the value of these two methods. Although our value assessments are conducted in predominately clear, variable, and cloudy conditions, the performance of the ensemble forecast methods varies with four primary cloud types: rare, transient, monsoonal, and convective. In Table 10, we replace the probability decision threshold with target priority indicating that a collection with a low (high) probability decision threshold is a high (low) priority target. Bold values identify the best results between the ensemble mean and weighted ranks methods. The grayed values indicate that no appreciable difference between the forecasts exists.

The weighted ranks method provides the most potential value for high priority targets (1–3) in rarely cloudy and transient cloud-cover conditions. Opportunities to collect clear imagery are numerous in these environments. Therefore, the tendency for the weighted ranks method to forecast clear conditions with high priority targets results in more clear collections than with the ensemble mean forecast. Although the weighted ranks method prefers clear, the ability of the forecast to correctly predict occasional cloud

cover is reflected in the expected utility value of 16%–25% increase in operational efficiency with high priority targets. Transient cloud-cover conditions such as clouds that accompany frontal systems are more challenging for the ensemble and the expected value for high priority targets is reduced to 8%–16%. This too, however, suggests that the addition information provided by the ensemble forecast has significant utility in collection operations.

The ensemble mean provides the most value for high priority targets in monsoonal and convective cloud-cover conditions. The moist bias of the ensemble mean reduces cloud filled collections. The risk-adverse user that has a high priority target can expected to gain a 10%–18% increase in operational efficiency in monsoonal cloud cover and 11%-19% increase in highly convective regions. The ensemble mean performs best in convective environments, but there is only a 1% difference between the expected value of the ensemble mean during monsoonal and convective cloud cover with high priority targets. Performance in porous cloud-cover conditions is similar to wide-spread cloud cover.

Furthermore, the differences between the ensemble mean and weighted ranks methods are not prominent. The mean difference between in expected value of the two forecasts for high priority targets is 2%—the mean is taken for all cloud cover types. The mean difference for medium priority targets (4–6) is less than 1%. The differences are unimpressive and suggest that neither ensemble forecast method can truly be expected to provide superior value over another. All, however, demonstrate the potential to add significant value to cloud-free collection operations.

Table 10. Ensemble mean and weighted ranks expected utility value calculations relative to type of cloud cover. Expected value (%) for each forecast method is plotted for each target priority (**1–9** with **1** being the highest priority). **Bold** values indicate the largest utility value between the two methods and grayed values indicate similar values.

Cloud Cover Type		Target Priority								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Rare	EMean	14	19	23	26	28	31	**33**	**34**	**34**
	WRank	**16**	**21**	**25**	**28**	**30**	31	32	32	27
Transient	EMean	7	11	15	17	20	21	23	**25**	**26**
	WRank	**8**	**12**	**16**	**18**	20	21	23	22	22
Monsoonal	EMean	**10**	**15**	**18**	**21**	**24**	24	24	23	23
	WRank	4	11	16	20	23	**25**	24	23	23
Convective	EMean	**11**	**16**	19	22	25	27	**29**	30	**30**
	WRank	9	14	19	**23**	25	27	28	30	26

VIII. SUMMARY AND CONCLUSIONS

We develop an ensemble-based forecast system to explore the potential for forecasts including uncertainty information to provide value in support of space-based image collections. Others have used statistical techniques to produce probability cloud forecasts, but we use a dynamic, flow-dependent method of forecasting cloud cover uncertainty by introducing perturbations to the initial conditions and trajectories of a cloud advection model (ADVCLD). We combine AFWA's global cloud analysis and NCEP's global weather ensemble to produce the initial conditions of the ensemble. Algorithms from AFWA's cloud advection model are applied to each of the global weather ensemble-member forecasts to generate twenty separate cloud-free forecasts at approximately 24-km resolution. Limitations in the availability of GEFS data restrict our ensemble analyses of skill and value to a 12-month period.

Ensemble evaluations focus on three regions. The regions represent persistently clear (Region 1), cloudy (Region 3), and variable cloud cover (Region 2). These regions were also chosen because they are operationally significance to U.S. military land operations and reduce the possibility of foreshortening (image distortion due to the curvature of the earth). The regions also allow for skill and value assessments amid distinct seasonal changes in cloud cover.

The operationally relevant cloud-free forecast threshold (cloud cover less than 30%) is evaluated in the three climatologically different regions. Regional evaluations of skill suggest that day-night cloud classifications are all but transparent in WWMCA initialization and verification. However, persistent cloud cover information in WWMCA may produce a moist (dry) bias in the initialization and verification of the ensemble forecast. The analyses and forecasts favor cloud cover values near 0% and 100%, making skill metrics that assume normal statistics mostly inappropriate. Thus we focus on contingency table metrics at the 30% threshold.

We use the Heidke Skill Score (HSS), True Skill Score (TSS), and Odds Ratio Skill Score (ORSS) to assess skill, or accuracy, of the ensemble in discriminating between clear and cloudy events. Although significant skill is demonstrated with respect to the HSS and TSS, skill information provided by the ORSS directly answers one of the decision-maker's primary questions; how well does the forecast correctly predict clear conditions versus miss opportunities to collect clear imagery? More importantly, how useful is the forecast information to the decision maker?

Utility theory offers a convenient way to quantify the value of forecasts in mitigating losses associated with uncertain events. Utility theory facilitates the consideration of forecast value without the explicit use of economic information. Rational users can assign utility values to potential forecast outcomes. These values can then be used to calculate a user's optimal decision threshold. The relationship between the user's assigned utility values (sensitivity to cloud filled imagery) and probability decision threshold can be quite complex (Appendix A). In our first order forecast evaluation, the perceived utility of correctly rejecting cloudy imagery is synonymous with the user's probability decision threshold.

A. ENSEMBLE SKILL

We find that variations in the ability of the ensemble to distinguish between clear and cloudy events depend mainly on the type of cloud-cover challenges presented within each region. At the 20% probability decision threshold, where we suspect most high priority military cloud-free operations reside, differences in the mean skill of the ensemble mean and probability forecasts are apparent. The HSS and TSS are used to demonstrate the difference in skill from region to region. Although we did not further explore the apparent sub-regional skill, we note that seasonal changes in sections of a region can make noticeable differences in the regional skill of the ensemble. The overall skill within the regions is calculated using the ORSS, the most appropriate skill score for our problem. The ORSS allows us to determine which forecast is most likely to result in a clear collection versus missing a clear collection opportunity.

We find five primary consistencies that arise in our ORSS comparisons. 1) The ensemble mean performs best during rarely cloudy events, April–June in Region 1. 2) The weighted ranks method performs best during transient cloud cover events, February–March in Regions 1 and February–April in Region 2. 3) Appreciable differences do not exist during widespread, monsoonal cloud cover, Jun–August in Regions 1 and 2. 4) The ensemble mean frequently performs best in cloud cover environments that are characterized by widespread convection, as in Region 3. 5) The ensemble control is not as skillful in discriminating between clear and cloudy events in extremely variable and convective, or highly uncertain, cloud-cover environments as in Regions 2 and 3.

B. ENSEMBLE UTILITY

In the absence of economic information, we use utility theory to assess the expected value of ensemble forecasts for cloud-free collection operations. Economic value is often cited in research but rarely used in practice because military decision makers cannot or often do not quantify the cost/loss ratio appropriate for their operations. We find utility value to be a more practical approach to obtaining forecast value. Therefore, we examine the expected utility of two types of users; risk-neutral and risk-averse. Risk lovers are assumed to be anomalous among military decision makers.

Utility theory allows us to calculate and compare the expected utility value of our ensemble forecast methods. We evaluate the expected utility value at nine probability decision thresholds (target priorities) for risk-neutral and risk-averse users. Risk-neutral users experience larger changes in expected value across the range of probability decision thresholds than risk-averse users. Forecasts at the 20% probability decision threshold are not always useful input to the decision process of risk-neutral users. Independent of region or season, risk-neutral users do not attain appreciable benefit from including the forecast in collection planning processes. However, military operations are rarely indifferent to risk.

When a user is risk-averse, forecasts can potentially add value at all decision thresholds. Low-priority targets receive the most potential improvement in efficiency,

but these targets tend to be targets of convenience and are not highlighted in the context of this research. The potential for risk-averse users to gain appreciable value from forecasts with high priority targets is the most interesting result. Decision makers who desire to collect high priority targets can receive 4%–25% increase in collection efficiency by using the forecasts.

The percent increase in collection efficiency depends more on cloud cover type than region. The average difference in expected value between the five ensemble forecast methods is about 1%–2%. Therefore, we cannot conclude that one ensemble forecast method should be preferred to another. We do, however, find that the biases of each method are reflected in the comparative performance within each cloud cover regime. The weighted ranks provides the most expected value in mostly cloud-free conditions and the ensemble mean performs best in cloud-filled conditions. The weighted ranks method produces the largest expected value for rare (16%–25%) and transient (8%–16%) cloud cover events with respect to high priority targets. The ensemble mean forecast produces the most expected utility value for high priority targets in monsoonal and convective cloud cover conditions.

C. LIMITATIONS

Using WWMCA for both initialization and verification of the ensemble forecasts is less than optimal. The impact of persistent cloud data, which remain in the absence of newer satellite data, is not examined, but we suspect that lags exist in cloud advection, development, and destruction as compared to reality. Global calculations of WWMCA spatial variance and autocorrelation confirm that using WWMCA as input for model initiation as well as verification is not expected to produce erroneous verification results, but a uniquely different verification source with error statistics smaller or similar to WWMCA is preferred to ensure forecast and analysis autonomy. Unfortunately, no other global cloud observation system exists, which provides data at regular intervals and at a comparable spatial resolution.

Skill and value assessments are made based on a 12-month dataset rather than climatology. Cloud-cover frequency is predominately used when describing the probability of cloud cover based on the analysis (e.g., probability of clear) and climatology is used when describing the forecast probability (e.g., climatological forecast). Although we show evidence that the probabilities herein can reasonably represent the true proportion of cloud cover in each region, using the terms frequency and climatology interchangeably can be misleading. Judging forecast based on the frequency of cloud cover in a given month is practical, but our limited data set does not allow us to conclude that the forecasts will most often perform this way. A true climatology is required for this assertion.

One of the primary factors in calculating expected utility value is the probability of clear conditions. Currently, we have no means to determine if the frequency of cloud cover used to calculate the expected utility value of the forecasts truly represents the climatology of the region. Monthly data that spans several years is preferred to calculate true climatological probabilities and subsequently the utility inherent to the operation as defined by climatology.

The bi-modal distribution of the ensemble and analysis discourages the evaluation of other cloud-fraction thresholds. Other reasonable thresholds (e.g., 20% and 40%) are assumed to yield similar results per the tendency of the ensemble and analysis to favor 0 and 100% cloud cover. This distribution highlights the insufficient diversity in the ensemble-member forecasts, which limits the application of forecast uncertainty in collection operations.

The cloud-forecast models used to build the ensemble are antiquated compared to current system capabilities. The 6 h temporal resolution undoubtedly smoothes or misses intermediate changes in cloud cover, which can have an impact on cloud cover advection. The spatial resolution of the ensemble is also a limiting factor. Producing a higher resolution ensemble could increase the number of grid points sampled in each region—if the correlation distance between grid points also decreases with sample size. Higher

spatial resolution does not equal better ensemble skill but moves ensemble evaluations closer to the resolution of current collection operations.

The advection scheme of ADVCLD is a poor representation of cloud cover processes and evolution. By advecting the total cloud fraction within the column, the cloud cover prediction does not account for differential advection between atmospheric layers. When the cloud cover is primarily convective, cloud development and decay is more important than advection, but the model only develops and decays clouds in the presence of ridges and troughs. The lack of complex physics in the development and decay of cloud cover significantly limits in the effectiveness of the ensemble forecasts in correctly predicting cloud cover fractions in dynamic regions.

D. RECOMMENDATIONS

Utility theory should be employed operationally in cloud-free forecast evaluations and expanded to other forecast applications. Employing expected utility value is a straightforward and practical method of capturing the potential value Air Force Weather forecasts have on military operations. Understanding where and when to apply resources benefits both decision makers and forecast providers. Informed decision makers can streamline their process and move from planning to execution without being overloaded with information that has little impact on the outcome of their operation. This same knowledge can be garnered by Air Force Weather to better manage weather resources. With the use of expected value information, forecast enhancements and support become justifiable, focused on the most critical weather requirements, and tailored to mitigate the risks associated with mission success.

More research is required before ensemble forecasting in support of space-based image collections can be made operational. Our research suggests that the ensemble demonstrates skill and can provide value to cloud-free operations. However, upgrades are needed to overcome the lack of spread between ensemble-member forecasts, coarse resolution of the ensemble, and the simple advection scheme.

Adjustments to the initialization algorithm could quite possibly improve the initial ensemble spread and skill of the forecast. Rather than choosing between GEFS and WWMCA values, their information can be combined. WWMCA provides the first guess analysis of the atmosphere and GEFS provides the uncertainty in the first guess. Both are important in developing reasonable expectations on how the atmosphere will develop. Two simple methods of combining the data would be to use a simple or weighted mean. By calculating the mean moisture between WWMCA and each GEFS perturbation, neither the analysis nor the model information is completely lost. This will eliminate the truncation problem of choosing between the two data sets, however, the ensemble spread will still be reduced significantly. A weighted mean would work better.

Applying weights to the datasets that prefer GEFS member forecasts to WWMCA values would also increase the ensemble spread over the current method. This method necessitates testing of various weight combinations to obtain the most optimal weighting factors. One weighting option is 1:0 (GEFS:WWMCA). Here only GEFS values are used and WWMCA values are eliminated from the 00-h forecast. This option allows for the maximum ensemble spread achievable relative to GEFS data. Although we have shown that WWMCA temporal correlations are small and suitable to be used as first guess cloud estimates of the initial forecast hour, a 1:0 weighting of the moisture fields facilitates the use of WWMCA as an independent analysis.

The lack of spread amongst the ensemble-member forecasts can perhaps be overcome by allowing the ensemble spread to define forecast uncertainty. Our research compares each ensemble member to the 30% cloud-free threshold, and the number of members forecasting clear versus the number forecasting cloud is used to define cloud-cover uncertainty. If all member forecasts are equally distributed with a minimum cloud-fraction value of 31% and maximum value of 100%, then the probability of cloudy conditions is 100%. This may misrepresents the true uncertainty in the ensemble cloud-cover prediction. Rather than requiring ensemble-member forecasts to cross the cloud-free thresholds to define probabilities less that 100%, the size of the ensemble spread can be used to define forecast uncertainty.

Two opportunities for increasing the ensemble resolution present themselves. First, the model can be rebuilt when NCEP makes the half-degree, 3-h GEFS forecasts operational. This method increases the resolution of the cloud-free forecast ensemble and perhaps may better reflect the degree of uncertainty in the forecast. Another option is to replace the underlying global advection model with a model that parameterizes cloud formation, maintenance, and dissipation (Jakob 2003). This also facilitates the use of WWMCA as an independent verification tool.

Cloud-scale processes are typically too small and complex for typical atmospheric models to resolve. Cloud development, maintenance, and decay depend on atmospheric location (e.g., vertical, meridional, surface), atmospheric constituents (e.g., aerosols, water vapor), and physical processes (e.g., turbulence, radiation, cloud microphysics). One way to ensemble current model cloud process is to parameterize the uncertainty in these processes. Adding uncertain in cloud development, maintenance, and decay to the uncertainty in cloud movement will undoubtedly make for a better probabilistic cloud-cover prediction.

Our examination of five ensemble forecast methods leads us to three primary conclusions. 1) The current GEFS configuration combined with AFWA's ADVCLD algorithm produces a cloud ensemble forecast that lacks the physical complexity in cloud development, maintenance, and decay and has insufficient temporal and spatial resolutions to demonstrate substantial differences in skill and value between the ensemble mean, control and probability forecasts. 2) Appreciable skill and value exists in cloud-free forecasts and skill and value varies with cloud type and frequency. 3) Using utility theory to calculate the value of predictions is a viable means to measure the usefulness of forecasts to the military operations they support.

LIST OF REFERENCES

Anderson, J. L., 1996: A method of producing probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530.

Arkin, A., 1996: Absorption of solar energy in the atmosphere: Discrepancy between model and observations. *Science*, **273**, 779–782.

Audi, R., 1999: *The Cambridge Dictionary of Philosophy*. Cambridge: Cambridge University Press, [Available online at http://www.credoreference.com.libproxy.nps.edu/book/cupdphil#C_2.]

Barry, R. G., and R. J. Chorley, 1998: *Atmosphere, Weather and Climate*, 7th ed. Routledge, 409 pp.

Beach, L. R., and T. Connolly, 2005: Subjective probability and utility. *The Psychology of Decision Making: People in Organizations*, A. Bruckner and D. Breti, Eds., Sage Publications, 63–78.

Bentham, J., 1823: *An introduction to the principles of morals and legislation*. Clarendon Press, Utility Theory: A Book of Readings, John Wiley & Sons, 3–29.

Bishop, C. H., and Z. Toth, 1999: Ensemble transformation and adaptive observations. *J. Atmos. Sci*., **56**, 1748–1765.

Chou J., 1989: Predictability of the atmosphere, *Adv. Atmos. Sci.,* **6** (3), 335–346

The Centre for Australian Weather and Climate Research, cited 2010: Forecast Verification: Issues, Methods and FAQ. [Available online at [http://www.cawcr.gov.au/projects/verification/.]

A Dictionary of Philosophy, Macmillan, 2002: A Dictionary of Philosophy, Macmillan, A. Flew and S. Priest, Ed., Macmillan Publishers, 433 pp. [Retrieved from http://libproxy.nps.edu/form?qurl=http%3A%2F%2Fwww.credoreference.com/entry/macdphil/a_dictionary_of_philosophy_macmillan.]

Davidson, P., 1991: Is probability theory relevant for Uncertainty? A post Keynesian perspective, *The Journal of Economic Perspectives,* **5** (1), 129–143.

Droegemeier, K. K., 1990: Toward a science of storm-scale prediction. Preprints, *16th Conf on Severe Local Storms*, Kananaskis Park, Alta., Canada, Amer. Meteor. Soc., 256–262.

Eckel, F. A., 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF Ensemble. M.S. thesis, Dept. of Meteorology, Air University, 133 pp.

Eckel, F. A., and M. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF Ensemble. *Wea. Forecasting*, **13**, 1132–1147.

Ehrendorfer, M., R. Errico, K. Raeder, 1999: Singular-Vector Perturbation Growth in a Primitive Equation Model with Moist Physics. *J. Atmos. Sci.*, **56**, 1627–1648.

Errico, R., R. Yang, M. Masutani, and J. Woollen, 2007: The use of an OSSE to estimate characteristics of analysis error. Meteorologische Zeitschrift (in press).

Essenwanger, O., and G. Haggard, 1961: The relationship between cloud cover and relative humidity. Final Rept. Project Order R-65-099856-SC-01-91, National Weather Records Center, Asheville, N.C. 78 pp.

EUMETSAT, cited 2007: Cloud detection for MSG-Algorithm theoretical basis document. [Available online at http://www.eumetsat.int/Home/Main/Access_to_Data/Meteosat_Meteorological_Products/Product_List/index.htm.]

Forecast Verification Glossary, updated 2007: [Available online at http://www.swpc.noaa.gov/forecast_verification/Glossary.html#G.]

Gillies, D., 2000: Introductory survey of the interpretations: some historical background, *Philosophical Theories of Probability.* Routledge, 1–14.

Hamill, T. M., 2000: Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Mon. Wea. Rev.*, **129**, 550–560.

Hoffman, R., and E. Kalnay, 1983: Lagged Average Forecasting, and Alternative to Monte Carlo Forecasting. *Tellus*, **35A**, 100–118.

Hogan, R. J., 2009: Verification of cloud-fraction forecasts. *Quart. J. Roy. Meteor. Soc.* **135**, 1494–1511.

Hogan, R. J., and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

Hogan, R. J., C. Jakob, and A. J. Illingworth, 2001: Comparison of ECMWF winter-season cloud fraction with radar-derived values. *J. Appl. Meteor.* **40**, 513–525.

Hou, D., 2010: A Stochastic Total Tendency Perturbation Scheme Representing Model-related Uncertainties in the NCEP Global Ensemble Forecast System, (to be submitted to *Tellus*): [Available online at http://www.emc.ncep.noaa.gov/gmb/yzhu/gif/pub/Manuscript_STTP_Tellus_A_HOU-1.pdf.]

HQ AFWA/2WXG/16WS, 2010: Algorithm description for the Cloud Depiction and Forecast System II. Air Force Weather Agency, 566 pp.

Jakob, C., and M. Miller 2004: Parameterization of physical processes: Clouds. *Encyclopedia of Atmospheric Sciences,* J. R. Holton, J. A. Pyle, and J. Curry, Eds., Academic Press, 1692–1698.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

Kiess, R. B., and W. M. Cox, 1988: The AFGWC automated real-time cloud analysis model. *AFGWC Technical Note* 88/001, Air Force Global Weather Central, 82 pp.

Kemp, E. M., and R. J. Alliss, 2007: Probabilistic cloud forecasting using logistic regression. Northrop Grumman Information Technology/TASC, **8A.8**, 7 pp.

Larence, D. B., 1999: *The Economic Value of Information*, Spring-Verlag, 393 pp.

Lee, W., 1971: *Decision Theory and Human Behavior*, John Wiley and Sons, 351 pp.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102**, 409–418.

Lindley, D.V., 1985: *Making Decisions*. 2nd ed. John Wiley and Sons, 207 pp.

Lorenz, E., 1975: Climate predictability: the physical basis of climate modeling. *WMO, GARP Publication Series*, **16**, 132–136.

Lorenz, E. N., 1982: Atmospheric Predictability Experiments with a Large Numerical Model. *Tellus*, **34**, 505–513.

——, 1969: The Predictability of a Flow Which Contains Many Scales of Motion. *Tellus*, **21**, 28–307.

——, 1993: *The Essence of Chaos*. University of Washington Press, 227 pp.

McDonald, D., 2011: Personal communication.

Mullen, S. L., and D. P. Baumhefner, 1988: The sensitivity of numerical simulations of explosive oceanic cyclogenesis to changes in physical parameterizations. *Mon. Wea. Rev.*, **116**, 2289–2339.

Murphy, A. H., 1993: What is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Amer. Meteor. Soc.*, **8**, 281–293

——, 1966: A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio situation. *J. Appl. Meteor.*, **5**, 534–537.

Murphy, A. H., and M. Ehrendorfer, 1987: On the relationship between the accuracy and value of forecasts in the cost-loss ratio situation. *Wea. Forecasting*, **2**, 243–251.

NASA, cited 2011: Remote Sensing, [Available online at http://earthobservatory.nasa.gov/Features/RemoteSensing/remote_04.php.]

North, D.W., 1968: A Tutorial Introduction to Decision Theory. *IEEE Transactions on Systems Science and Cybernetics,* **SSC-4** (3), 200–210.

Palmer, T., and R. Hagedorn, 2006: *Predictability of weather and climate*, Cambridge University Press, 702 pp.

Press, W. H., and S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran.* 2nd ed. Cambridge University Press, 963 pp.

Roach, W.T., 1994: Back to Basics: Fog: Part 1 Definitions and basic physics. *Weather,* **49**, 411–415.

Raftery, A. E., F. Balabdaoui, T. Gneiting, and M. Polakowski, 2003: Using Bayesian Model Averaging to Calibrate Forecast ensembles. *Technical Report*, No. 440

Shorr, B., 1966: The cost/loss utility ratio. *J. Appl. Meteor.*, **5,** 801–803.

Stephenson, D. B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting,* **15,** 221–232.

———, 2002: Glossary in "Forecast Verification," [Available online at http://www.bom.gov.au/climate/pi-cpp/training/nms/Glossary.pdf.]

Sivillo, J. K., J. E. Ahlquist, and Z. Toth, 1997: An ensemble forecasting primer. *Wea. Forecasting*, **12**, 809–818.

Tennekes, H., A. Bade and I. Opsteegh, 1986: Forecasting forecast skill. *Preceedings of the ECMWF Workshop on Predictability in the Medium and Extended Range, Reading, England.*

Thompson, J., 1952: On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Meteor. Soc.*, **33**, 223–226.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteorol. Soc.*, **74**, 2317–2330.

——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **127**, 3297–3318.

Versant Case Study, cited 2010: An Object Database for Large-Scale Simulations: Better Performance and More Powerful Algorithms. [Available online at http://www.versant.com/developer/resources/objectdatabase/Versant_large_scale_simulations.pdf.]

Wei, M., Z. Toth, R. Wobus, Y. Zhu, 2007: Initial perturbations based on the ensemble transform technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79.

———, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79.

Wild, M., and A. Ohmura, 1999: The role of clouds and the cloud-free atmosphere in the problem of underestimated absorption of solar radiation in GCM atmospheres. *Phys. Chem. Earth (B)*, **24** (3), 261–268.

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Appl. Meteor.,* **8**, 209–219.

———, 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

Weymouth, G.T., T. Boneh, P. Newham, J. Bally, R. Potts, A. Nicholson, K. Korb, 2007: Dealing with uncertainty in fog forecasting for major airports in Australia. *Proceedings of the Fourth International Conference on Fog, Fog Collection and Dew*, La Serena, Chile, Pontificia Universidad Catolica de Chile, 73–76.

Wobus, R., 2011: GEFS data distribution. [Available online at http://www.dtcenter.org/events/workshops11/det_11/Wobus_GEFS_products_Wobus.pdf.]

Yates, J. F., 1990: *Judgment and Decision Making,* Prentice Hall, 430 pp.

Zhu, Y., cited 2011: Upgrade of NCEP/GEFS and NAEFS. [Available at http://www.emc.ncep.noaa.gov/gmb/yzhu/.]

Zhu, Y., D. Hou, M. Wei, R. Wobus, J. Ma, B. Cui, and S. Moorthi, cited 2011: Next Global Ensemble Forecast System. [Available at http://www.emc.ncep.noaa.gov/gmb/yzhu/.]

APPENDIX A. MITIGATION UTILITY

Figure 29 is a gross oversimplification of the possible outcomes that can occur in cloud-free collection operations. So far we have treated image collection as a dichotomous decision process with four distinct outcomes, but the possibility of collecting substitute images is also important. When a satellite image passes over a region, an image will always be collected even if all prospective collections are expected to be cloudy. Therefore, operators who elect to abort an image collection must choose from an assortment of other images with unique factors that define the collection priory of the alternates with respect to the primary image. The complexity of this process is beyond the scope of this research, but we present a means of capturing the utility of collecting an alternate image.

"Good" outcomes, collecting clear images or not collecting cloudy images, naturally have what we call inherent utility. Cloudy forecasts cause decision makers to consider mitigation strategies, which can result in the collection of a clear alternate image. Therefore, we say that clear alternate images possess mitigation utility. "Bad" outcomes, collecting cloudy images or not collecting clear images, have no inherent utility, but the latter can have mitigation utility when a clear alternate image is collected. Hence, HTs and CRs have inherent utility, and MSs and CRs have mitigation utility which results from the possibility of collecting an alternate image.

An alternate image is defined based by its proximity to the primary image and the minimum probability of it being clear. Based on a hypothetical orbital altitude of 600 km and a max off-nadir viewing angle of 24 degrees (geometry approximated from GEOEYE-1 specifications), the alternate image chosen—the image with the highest probability of being clear—must be no more than ~270 km (tan(24) x 600km) from nadir. Assuming that the primary image is not at nadir (most often the case), we condition it to be at some distance between nadir and 135 km off-nadir. Therefore, the alternate image can occupy the maximum distance from the primary image (5 grid points) yet remain

within maximum off-nadir viewing distance of the sensor. All grid points within this distance are compared and the grid point with the minimum probability of being clear is used to define the maximum utility an alternate image can add to the inherent utility of an outcome.

When assigning utility values, we must assume that each primary image is more valuable than any alternate image considered and preventing the collection of a cloudy image (CR) is more significant than collecting an alternate image. Hence,

$$u(CR) > u(m)$$

where, $u(CR)$ is the utility of correct rejections and $u(m)$ is utility of mitigation. Next, we remain sensitive to the fact that the cumulative utility of the components must maintain the order of preferences (Assumption 1) and be greater than the inherent or mitigation utility alone (Assumption 3).

$$1 > u(CR) + u(m) > u(m) \qquad \begin{matrix} .4 \le u(m) < 1 \\ .0 \le u(m) \le .4 \end{matrix}$$

This restricts the $u(CR)$ to values between .4 and .9 and $u(m)$ to values between .0 and .4. Once the A decision maker who wishes to collect images in a region where the demand for targets is low may assign .8 to $u(CR)$ and 0 to $u(m)$. A decision maker who is faced with collecting imagery in a region where clear targets are sparse and difficult to collect may have a greater appreciation for alternate imagery and assign .5 to $u(CR)$ and .4 to $u(m)$.

Mitigation utility $u(m)$, received from collecting a clear alternate image can reduce the dissatisfaction of MSs. Missing the opportunity to collect a primary image is most often worse than collecting an alternate image. A situation could exist where a clear collection is neither required nor desired making the utility of the alternate zero relative to the utility of missing an opportunity to collect the primary image $u(MS)$.

$$u(m) \ge u(MS) \qquad u(MS) = 0$$

If we choose the median $u(m)$, .2, the range of acceptable $u(CR)$ values reduces to [.4, .8). If we then choose the median $u(CR)$ of these values as our base line value, the inherent utility of CRs becomes .6. Table 10 shows the cumulative utility which results from combining $u(CR)$, $u(MS)$, and $u(m)$. Utility of clear collections have the greatest utility. Utility of not collecting a cloudy image and possibility of collecting a clear alternate image has high utility but less than a clear collection. The positive utility in missing an opportunity to collect a clear image is wholly due to the possibility of collecting a clear alternate image.

Table 11. Utility value calculation based on collection decision and cloud condition.

	HT	FA	CR	MS
Collect Clear Image	1			
Collect Cloudy Image		0		
Not Collect Cloudy Image			.6	
Not Collect Clear Image				0
Collect Clear Alternate Image			.2	.2
Cumulative utility	1	0	.8	.2

In the background section, we demonstrated how economic value can be used to calculate a forecast value score. We also showed how expected utility,

$$\max_i \sum_{j=1}^{2} \overline{u}_i \qquad 37$$

expected utility of perfect information,

$$\sum_{j=1}^{2} \max_i \overline{u}_i \qquad 38$$

and expected utility of less than perfect information

$$\sum_{E} \max_i \sum_{j=1}^{2} p(E \mid q_j)\overline{u}_i \qquad 39$$

can be used to determine the expected value of perfect and imperfect information. In this section, we complete the equations by adding the mitigation component of the expected utility and calculate expected utility value.

The utility gained from the probability of collecting a clear alternate image depends on the probability that a clear alternate $p(ai)$ exists in the satellite's field of view. The inclusion of mitigation utility $u(m)$ produces the following equations,

$$\max_i \sum_{j=1}^{2} p_j[u_i + u(m)] \qquad u(m) = \begin{cases} 0, & i = 1 \\ u_{ai}\, p(ai), & i = 2 \end{cases} \qquad 40$$

expected utility of perfect information,

$$\sum_{j=1}^{2} \max_i p_j[u_i + u(m)] \qquad u(m) = \begin{cases} 0, & i = 1 \\ u_{ai}\, p(ai), & i = 2 \end{cases} \qquad 41$$

and expected utility of less than perfect information.

$$\sum_{E} \max_i \sum_{j=1}^{2} p(E \mid q_j)\, p_j[u_i + u(m)] \qquad u(m) = \begin{cases} 0, & i = 1 \\ u_{ai}\, p(ai), & i = 2 \end{cases} \qquad 42$$

Mitigation utility only impacts decision 2, not collect. Therefore, mitigation utility $u(m)$ equals zero when i equals one, when the decision is to collect.

The maximum expected utility and value that can be gained depends on the values assigned to both mitigation utility $u(m)$ and correct rejection utility $u(CR)$. From Figure 61, we see that expected utility increases with increasing probability of clear collection. The minimum expected utility of perfect information corresponds to the minimum utility value assigned to correction rejections $u(CR)$. The expected utility value's maximum peak occurs when the utility of correct rejections $u(CR)$ is at its maximum (.8) and decreases as the utility value of correct rejections decreases but not uniformly. The addition of $u(m)$ offsets the reduction in expected utility value below 50% probability of clear.

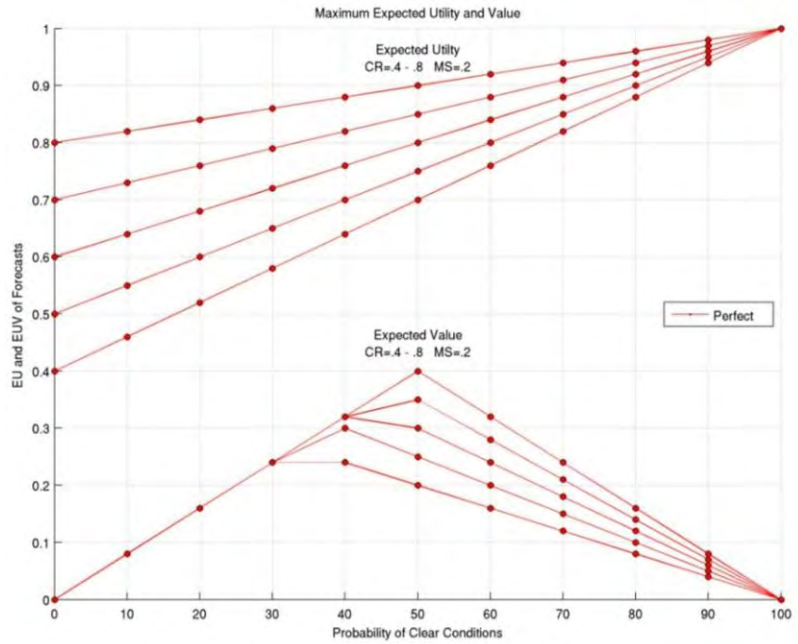


Figure 61. Variation in Correct Rejection Utility. The maximum achievable utility (upper) and value (lower) with additional information varies with the utility values assigned to Correct Rejections.

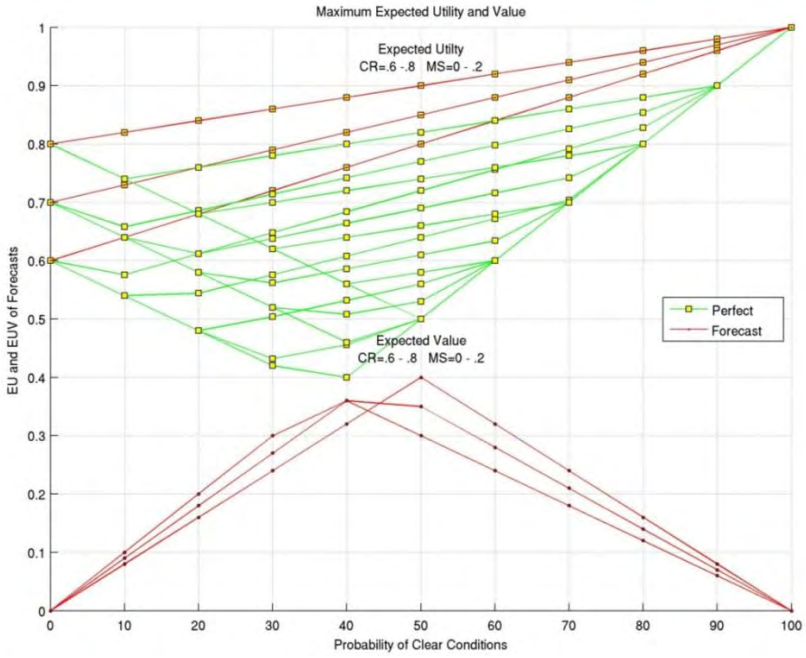
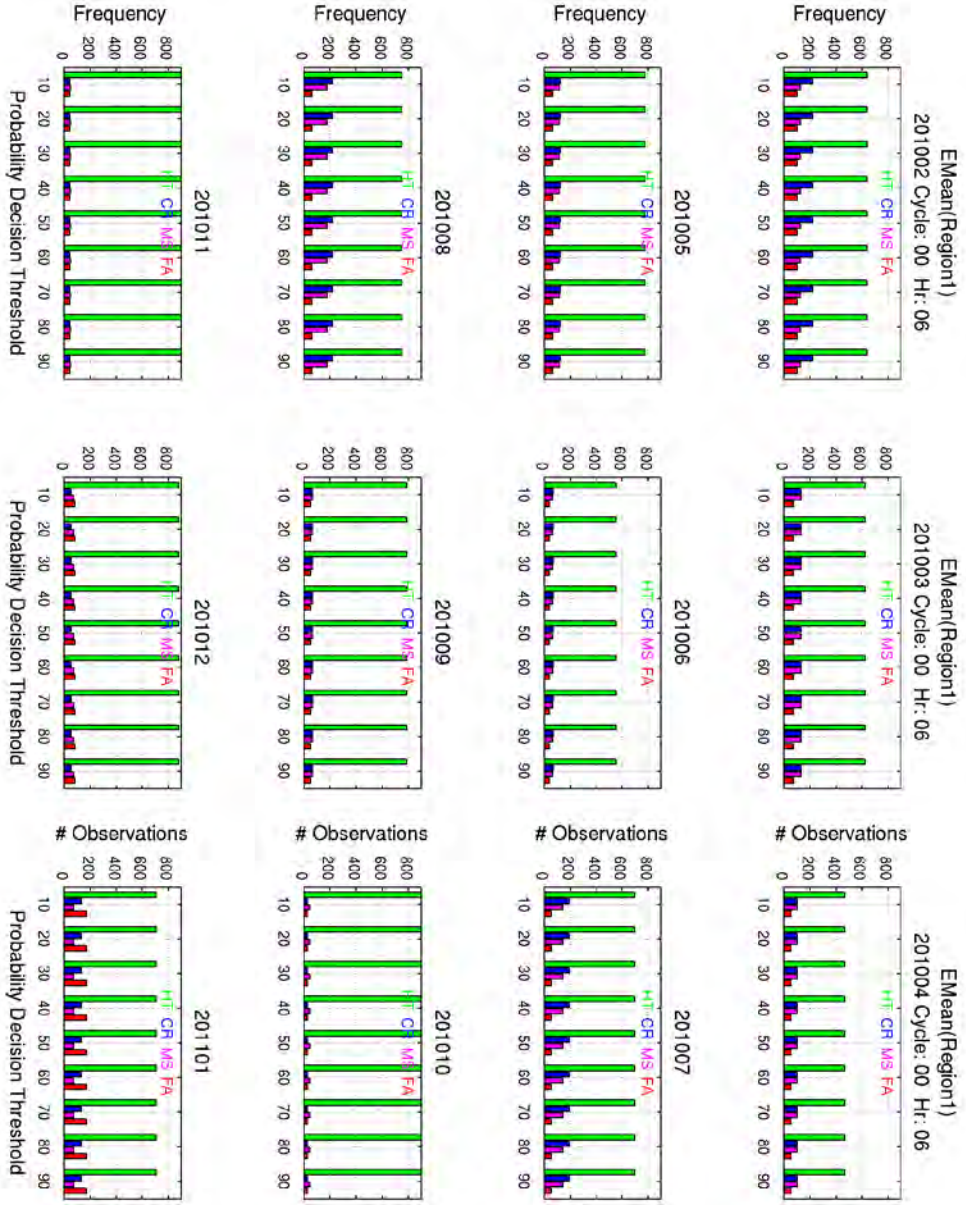
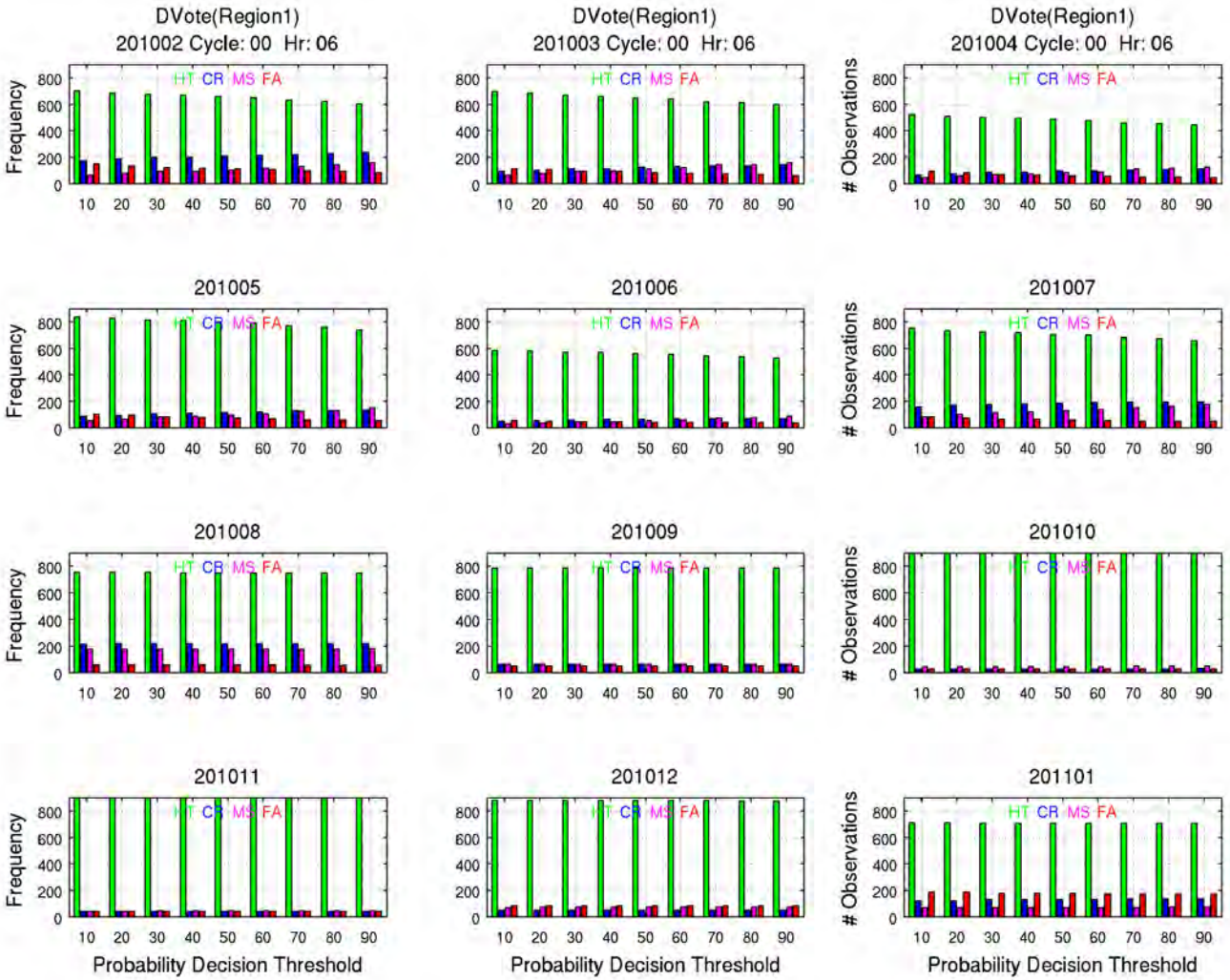


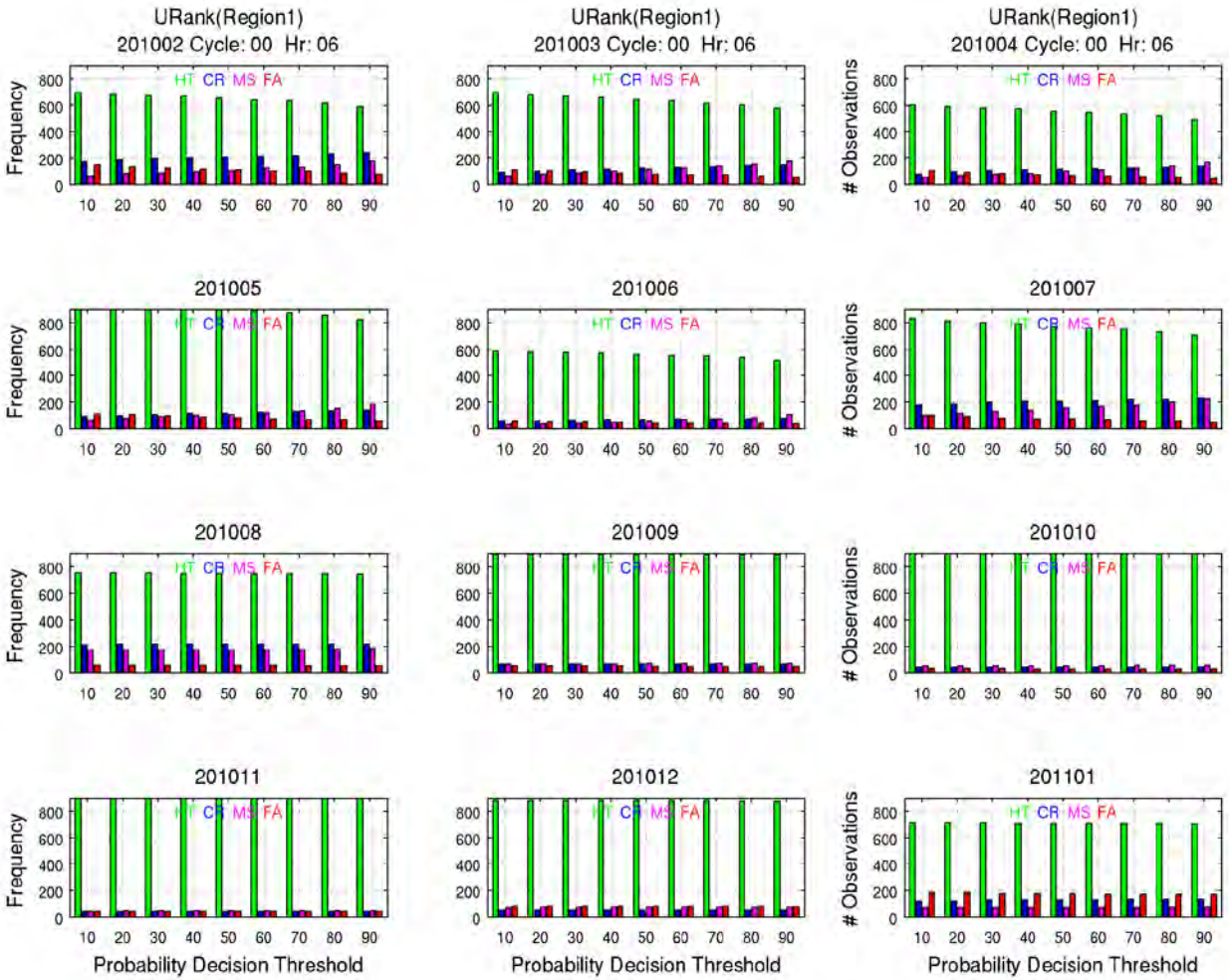
Figure 62. Variation in Mitigation Utility. The maximum achievable utility (upper) and value (lower) with additional information varies with the utility value assigned to alternate image collection

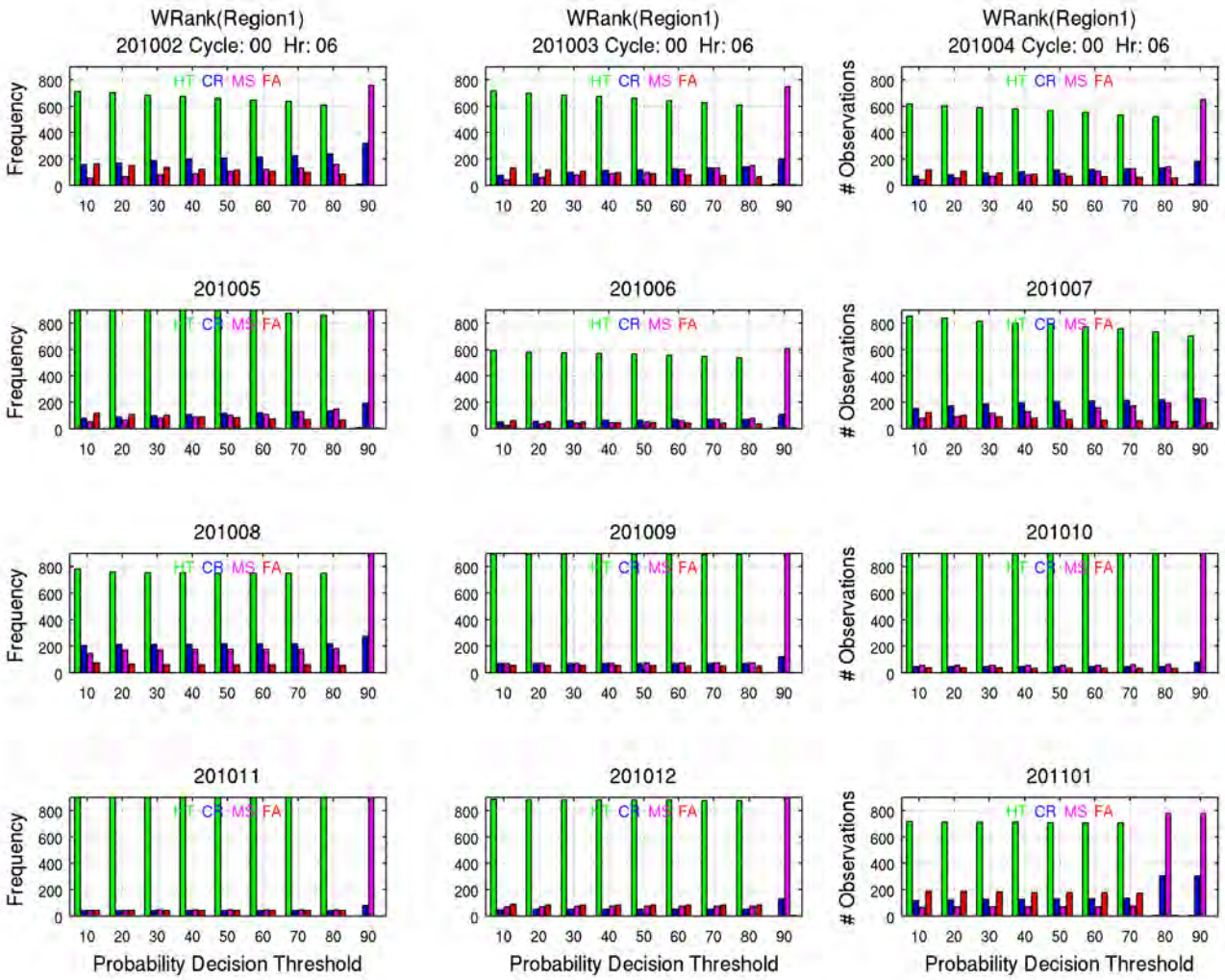
THIS PAGE INTENTIONALLY LEFT BLANK

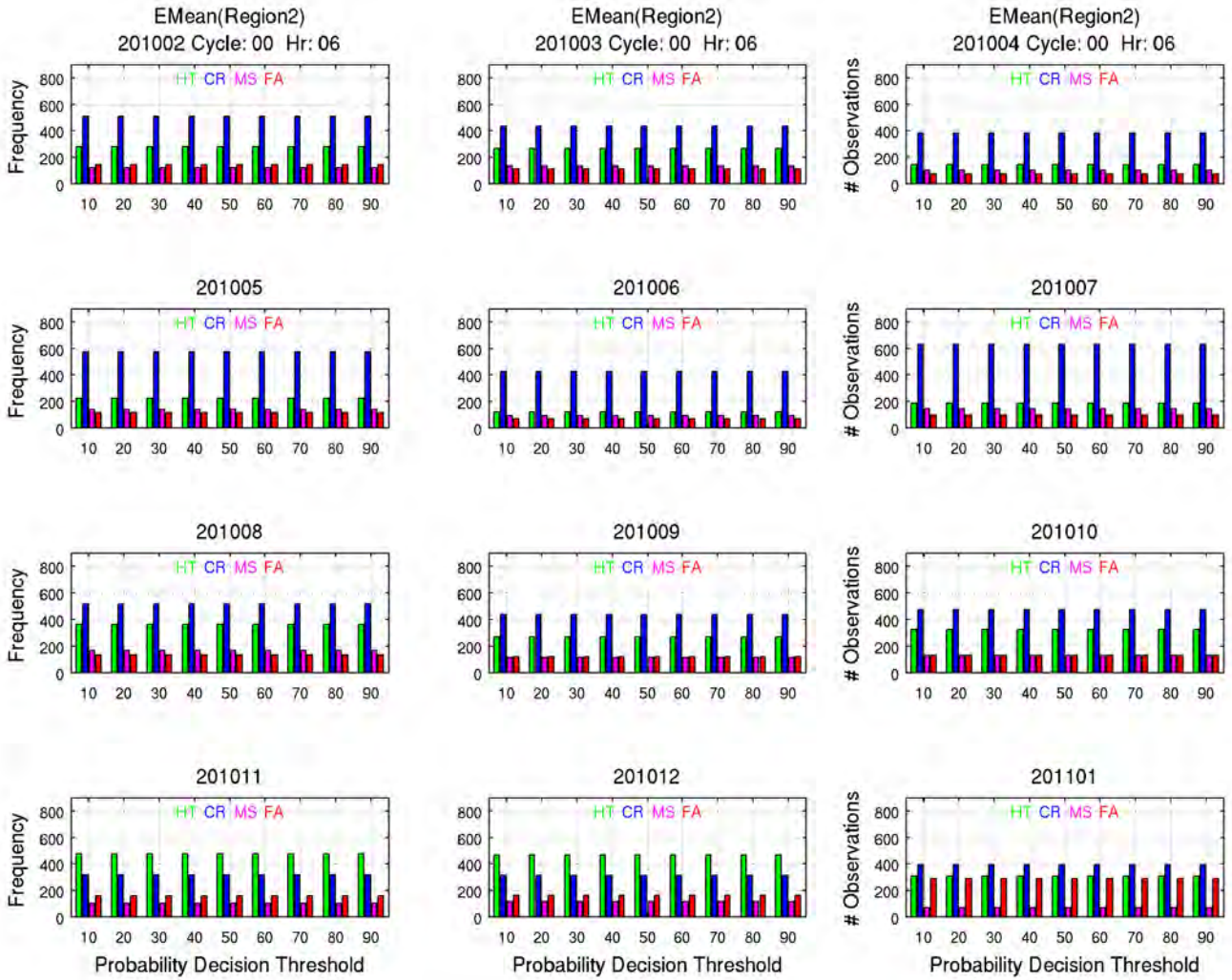
APPENDIX B. OUTCOMES AT PROBABILITY THRESHOLDS

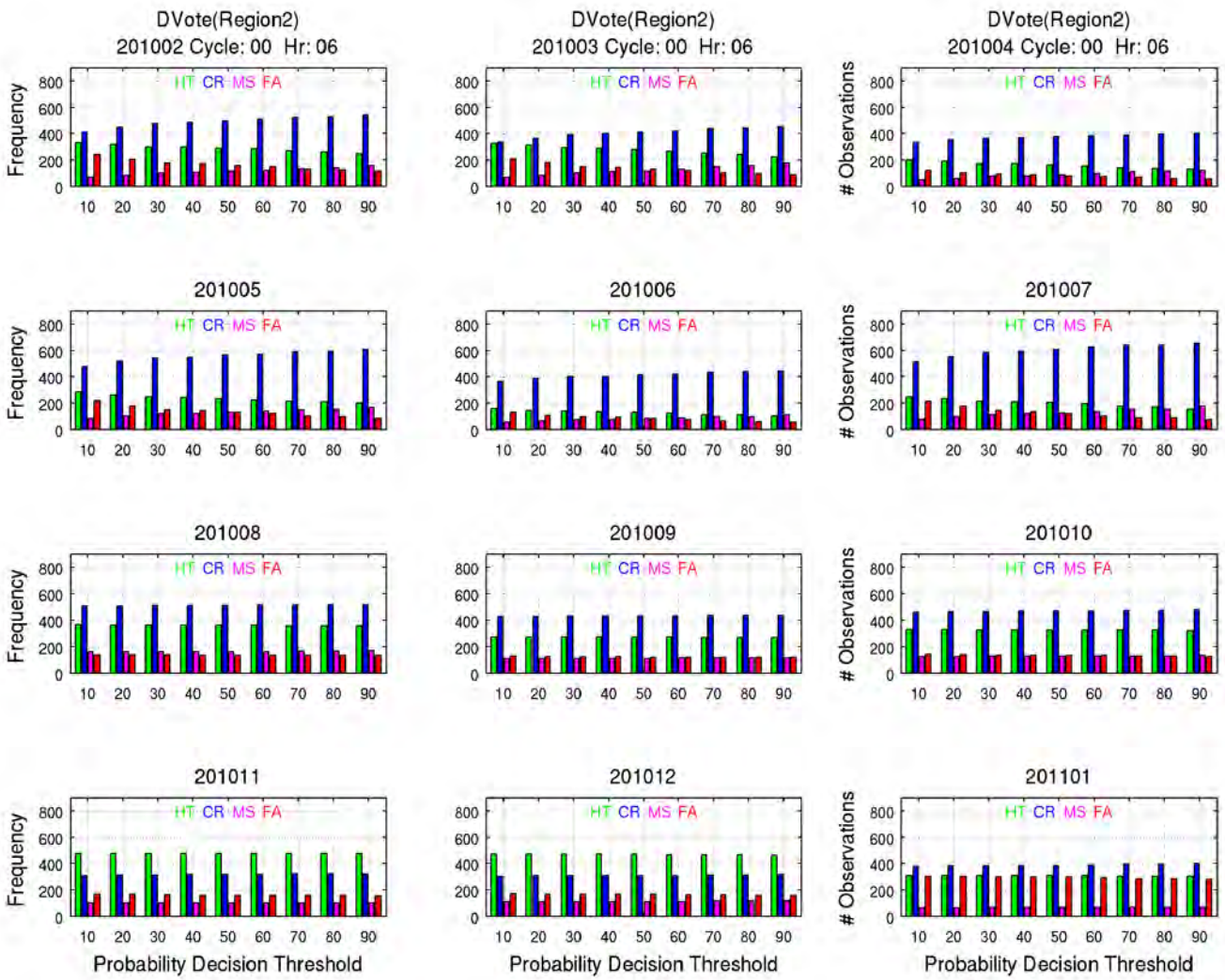


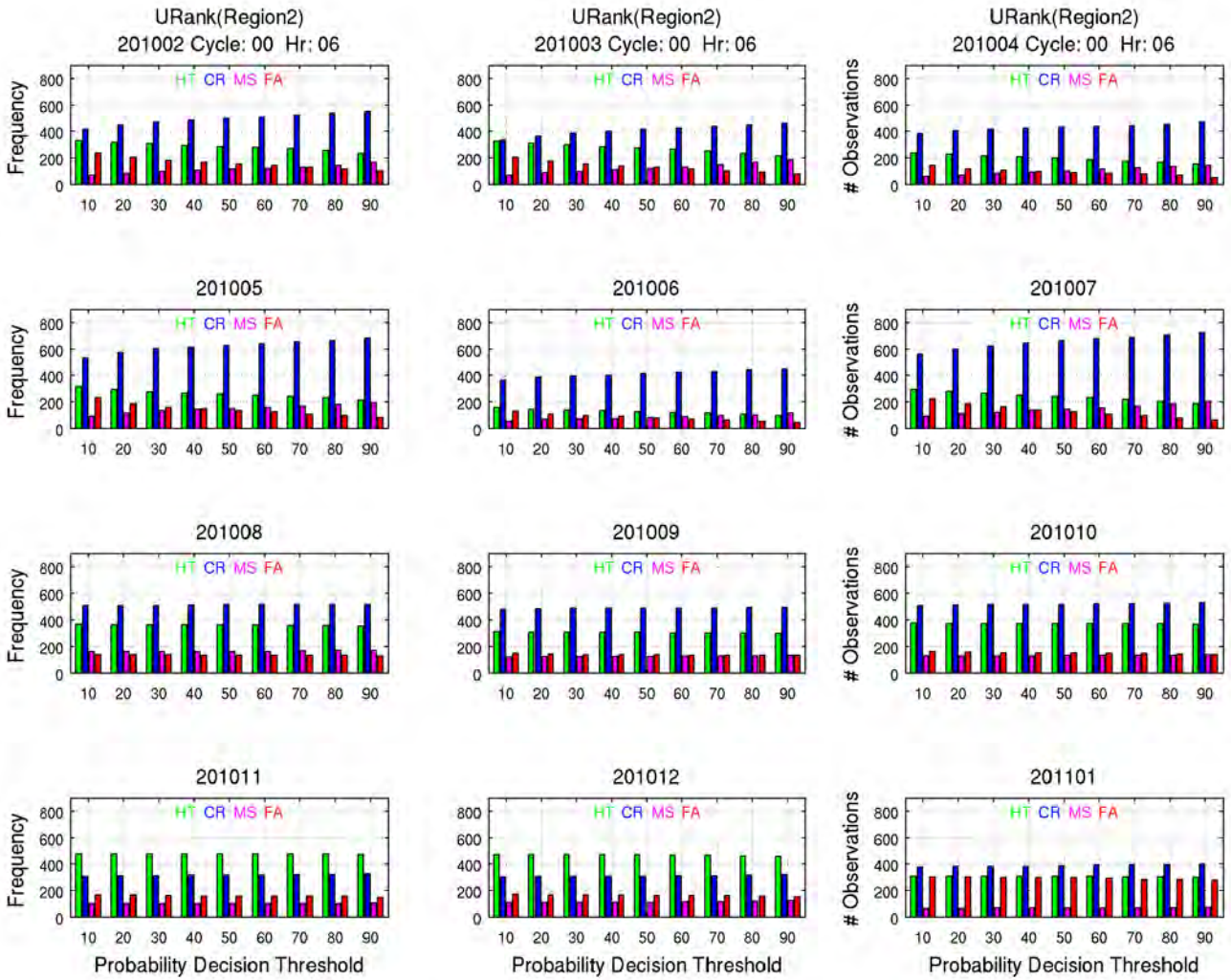

DVote(Region1)
201002 Cycle: 00 Hr: 06
DVote(Region1)
201003 Cycle: 00 Hr: 06
DVote(Region1)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold

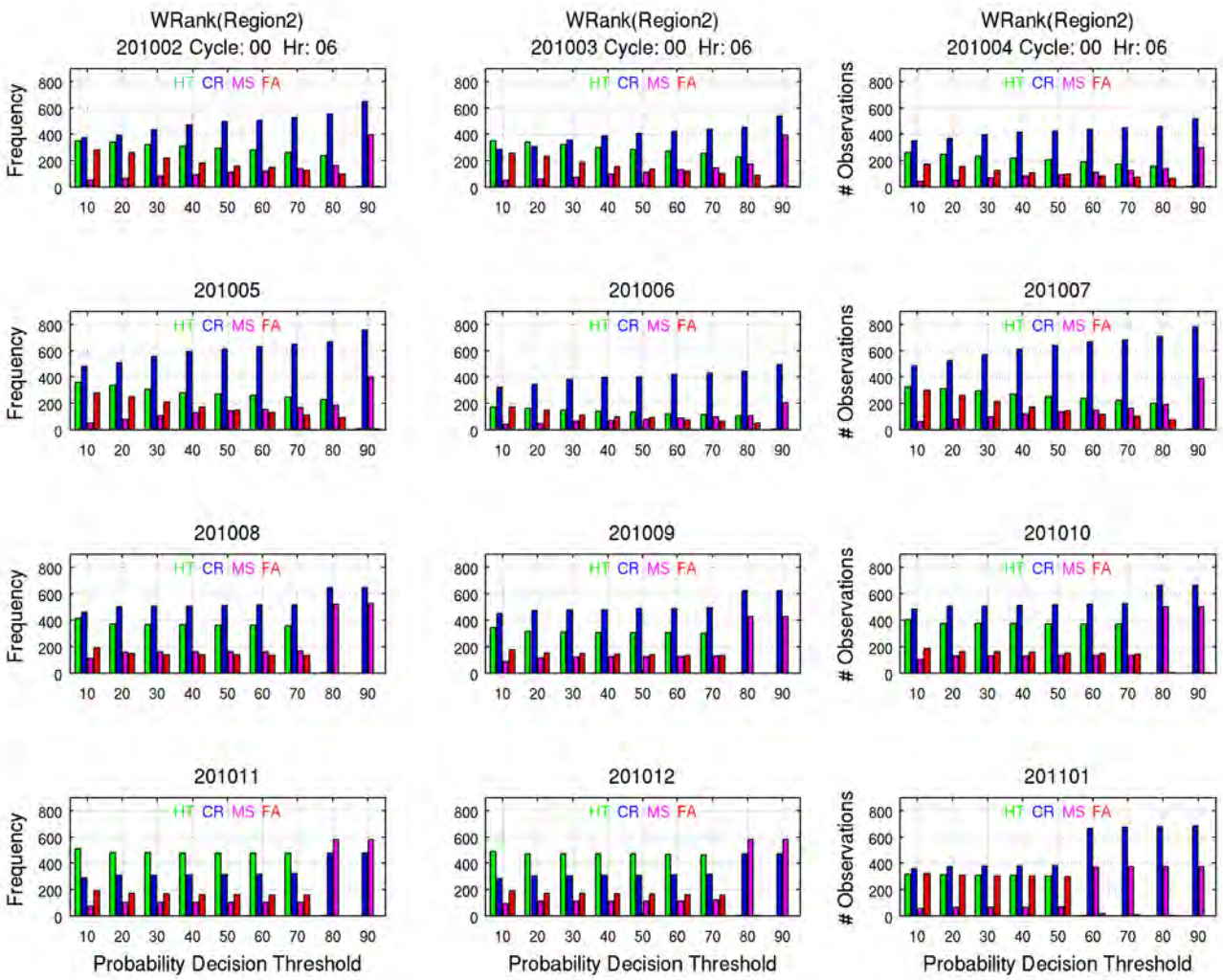

URank(Region1)
201002 Cycle: 00 Hr: 06
URank(Region1)
201003 Cycle: 00 Hr: 06
URank(Region1)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold

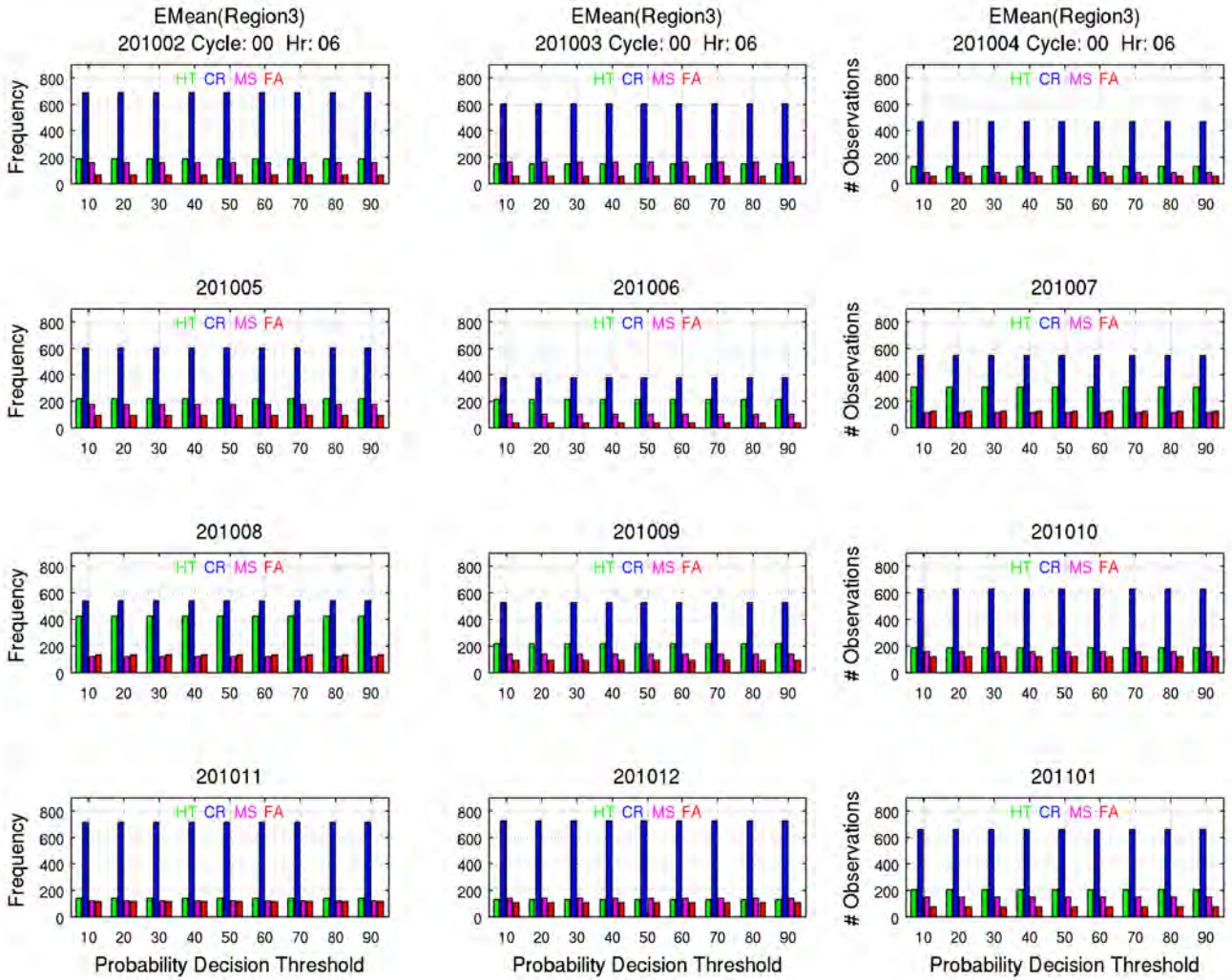

WRank(Region1)
201002 Cycle: 00 Hr: 06
WRank(Region1)
201003 Cycle: 00 Hr: 06
WRank(Region1)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold

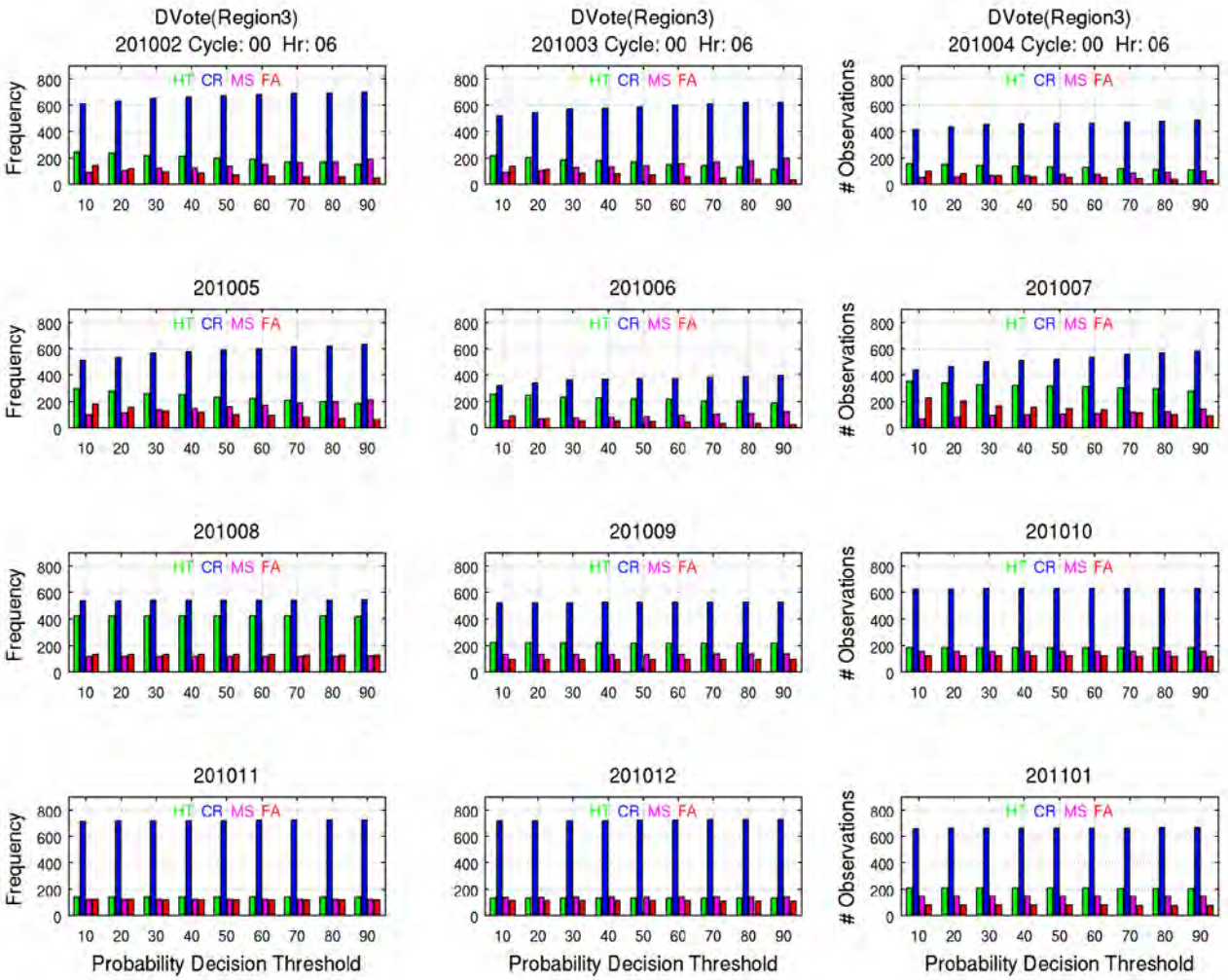

EMean(Region2)
201002 Cycle: 00 Hr: 06
EMean(Region2)
201003 Cycle: 00 Hr: 06
EMean(Region2)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0
200
400
600
800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold

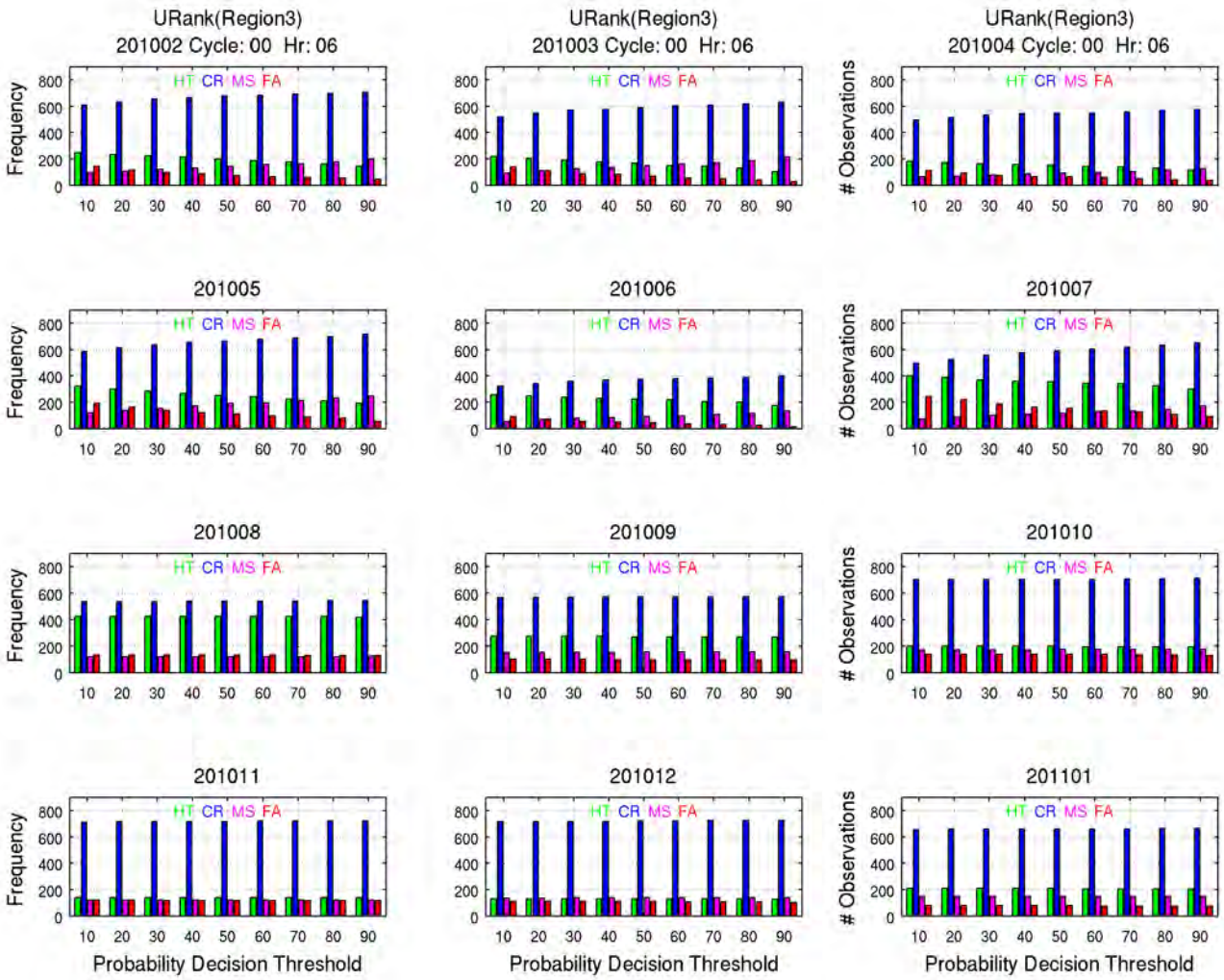

DVote(Region2)
201002 Cycle: 00 Hr: 06
DVote(Region2)
201003 Cycle: 00 Hr: 06
DVote(Region2)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold

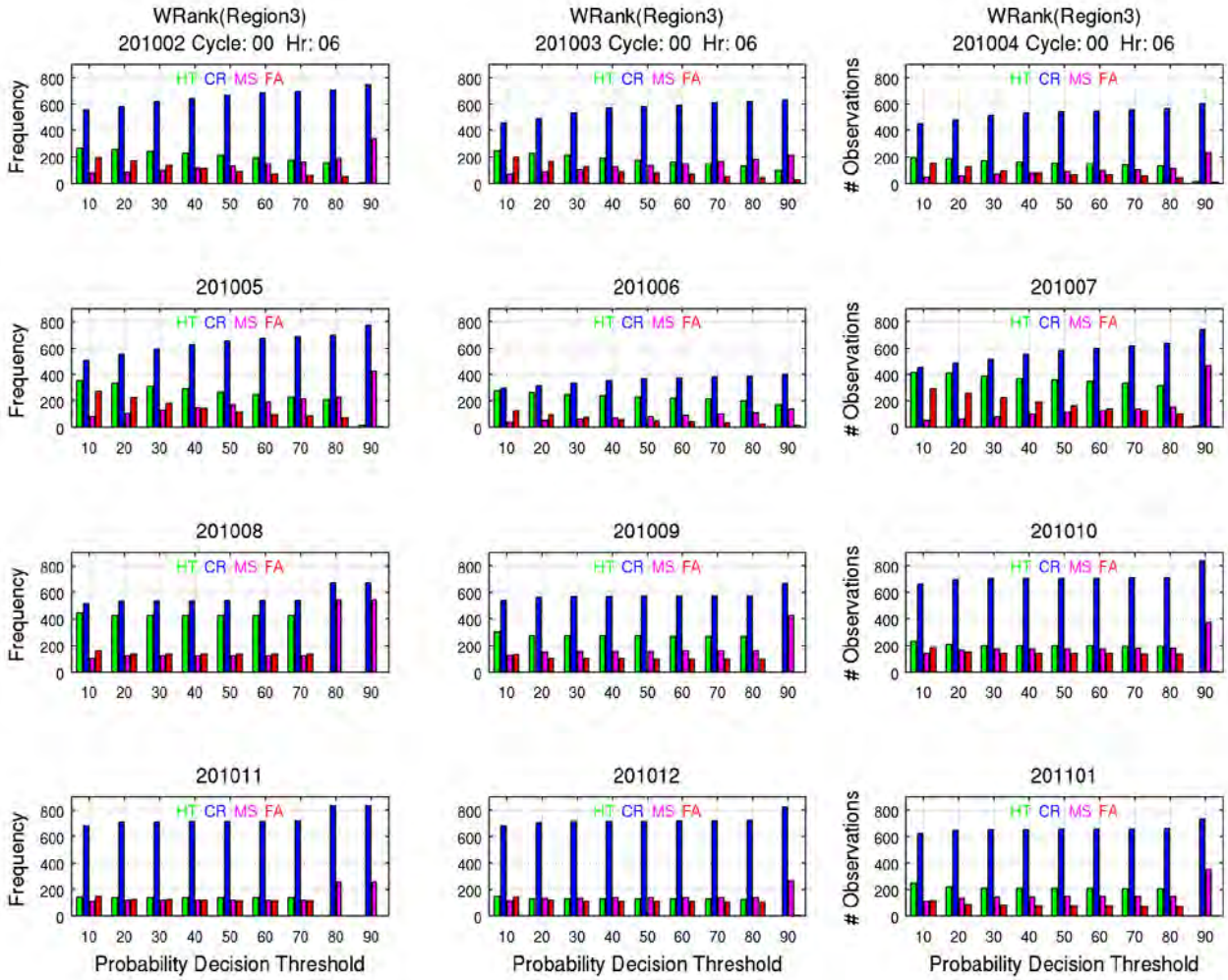

URank(Region2)
201002 Cycle: 00 Hr: 06
URank(Region2)
201003 Cycle: 00 Hr: 06
URank(Region2)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
Probability Decision Threshold
0
200
400
600
800
10 20 30 40 50 60 70 80 90


WRank(Region2)
201002 Cycle: 00 Hr: 06
WRank(Region2)
201003 Cycle: 00 Hr: 06
WRank(Region2)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold

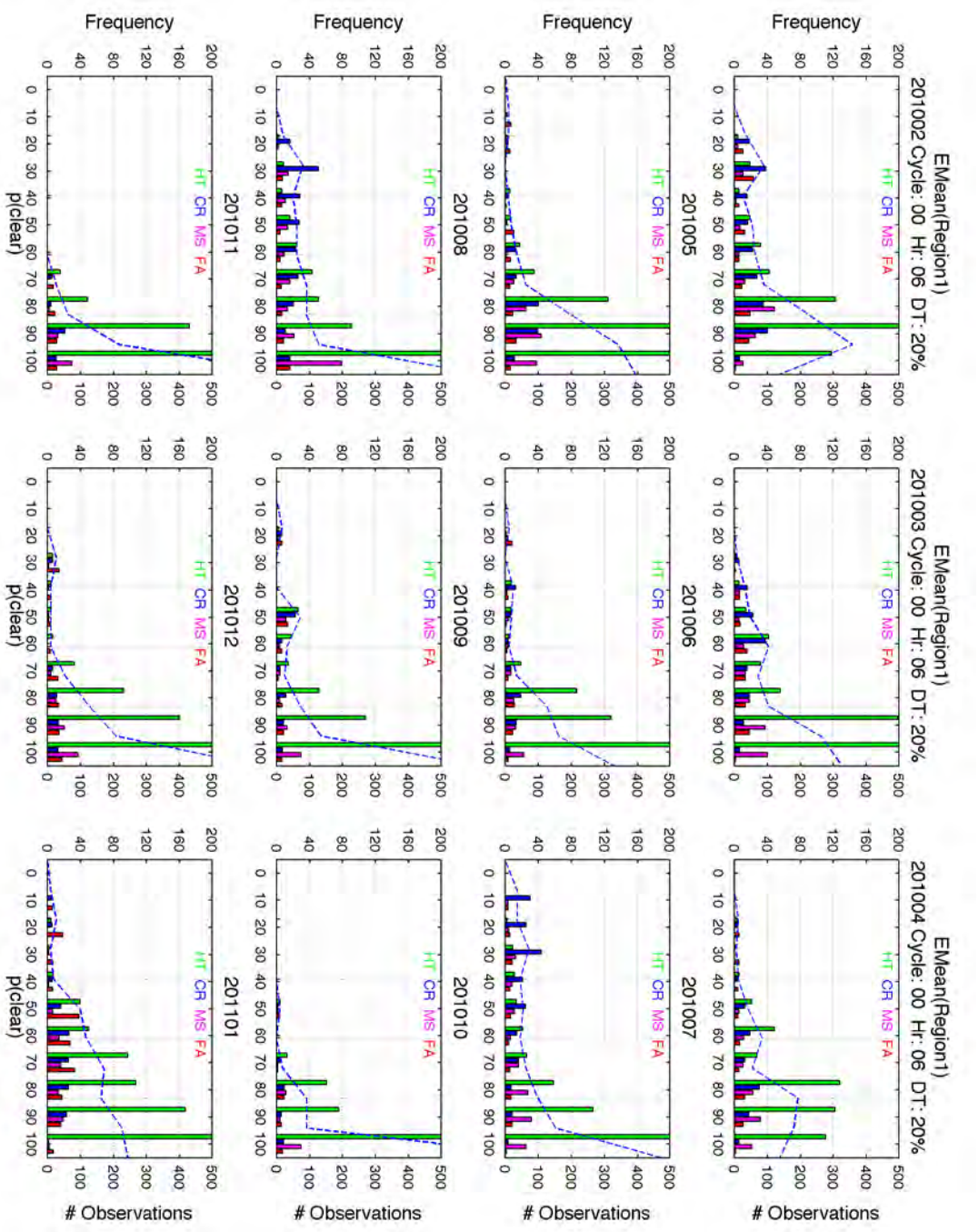

EMean(Region3)
201002 Cycle: 00 Hr: 06
EMean(Region3)
201003 Cycle: 00 Hr: 06
EMean(Region3)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold


DVote(Region3)
201002 Cycle: 00 Hr: 06
DVote(Region3)
201003 Cycle: 00 Hr: 06
DVote(Region3)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold


URank(Region3)
201002 Cycle: 00 Hr: 06
URank(Region3)
201003 Cycle: 00 Hr: 06
URank(Region3)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold

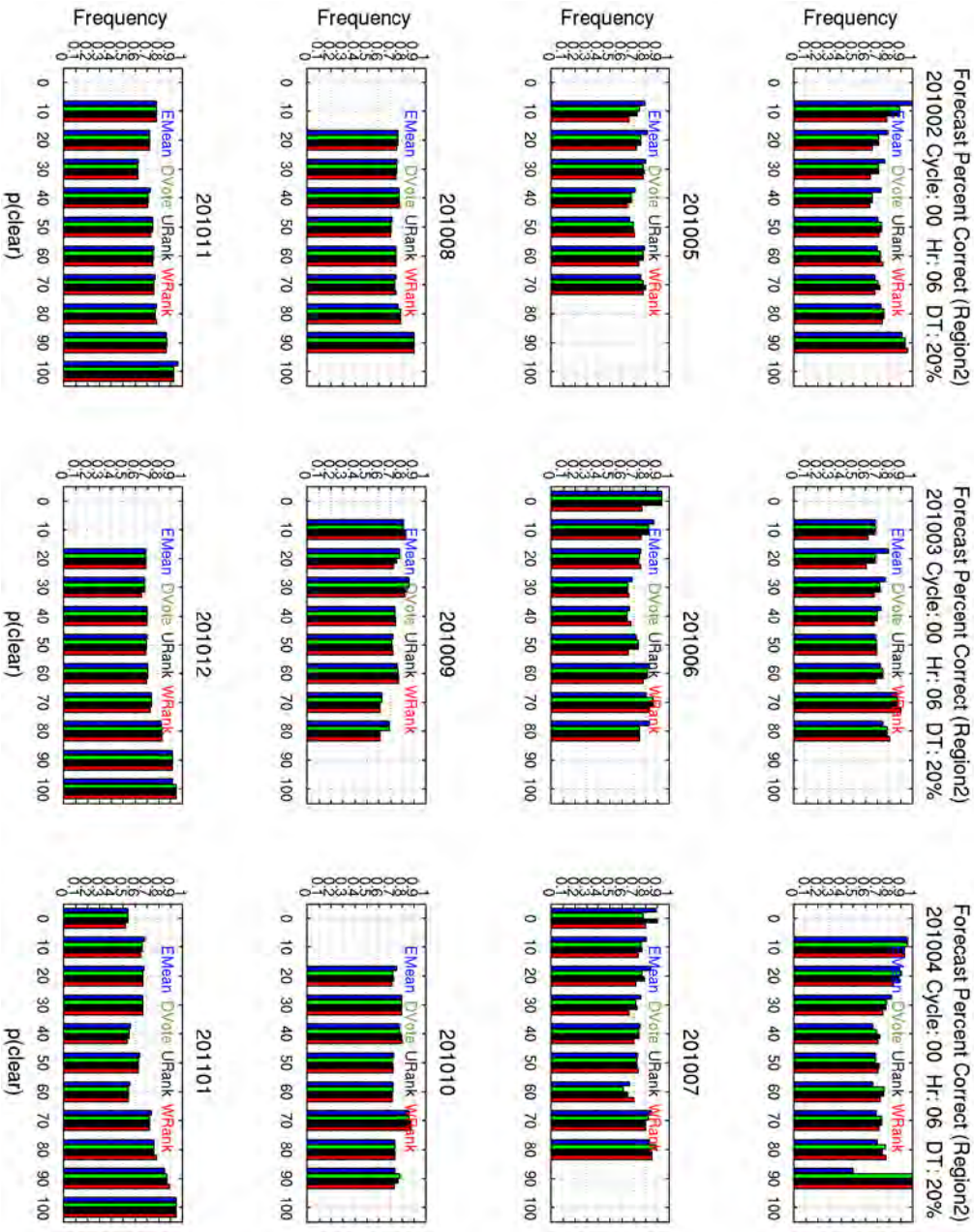

WRank(Region3)
201002 Cycle: 00 Hr: 06
WRank(Region3)
201003 Cycle: 00 Hr: 06
WRank(Region3)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
HT CR MS FA
Frequency
Observations
0 200 400 600 800
10 20 30 40 50 60 70 80 90
Probability Decision Threshold

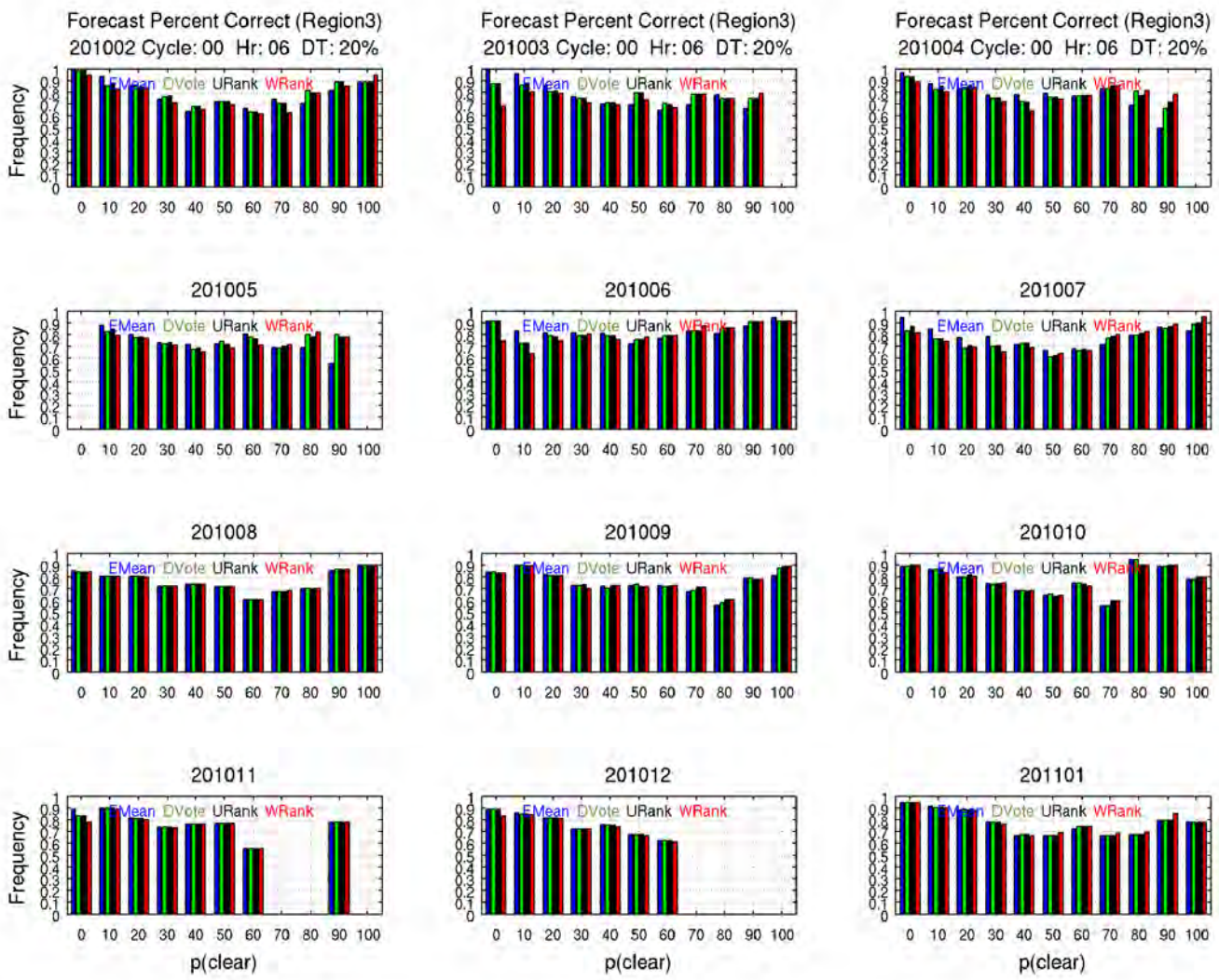
APPENDIX C. OUTCOMES AT FREQUENCY ≤ 30%

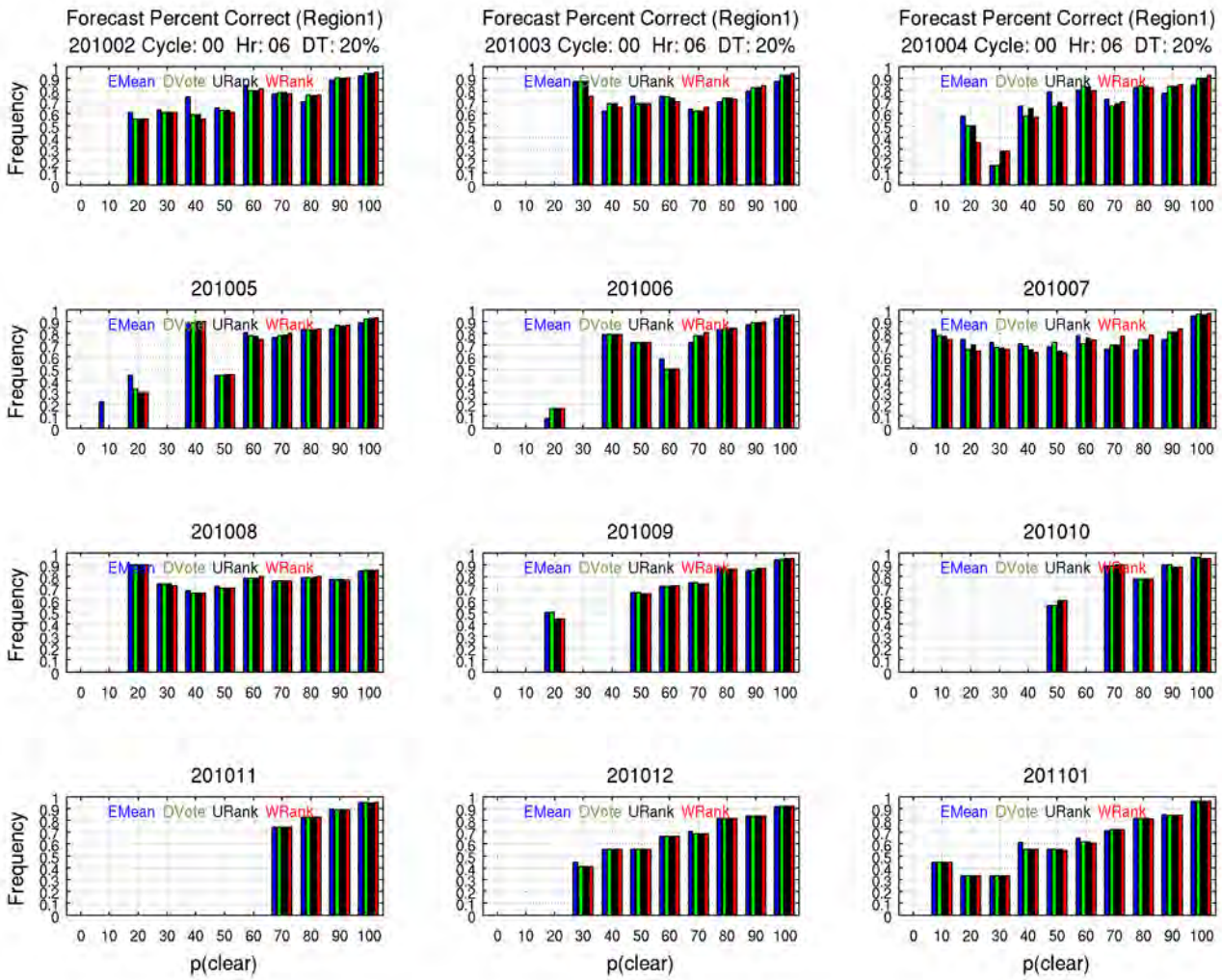


THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. PERCENT CORRECT PLOTS

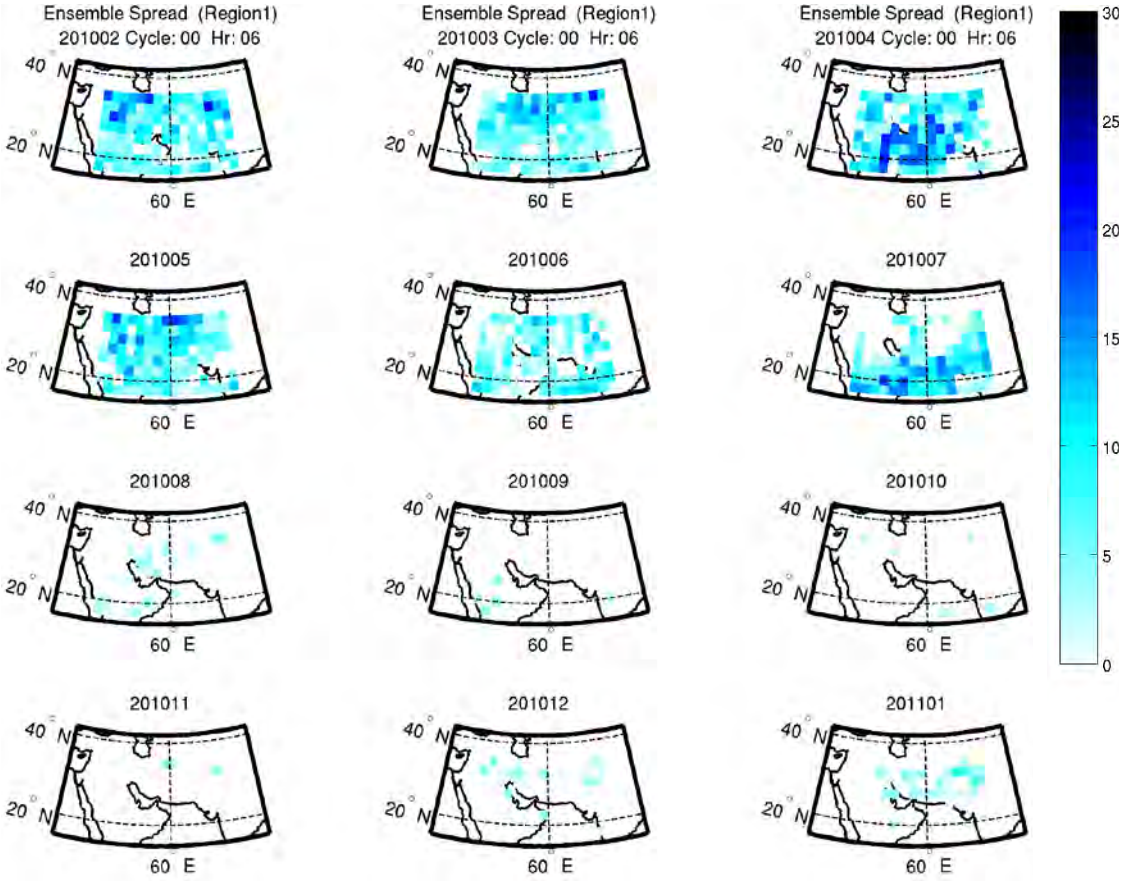


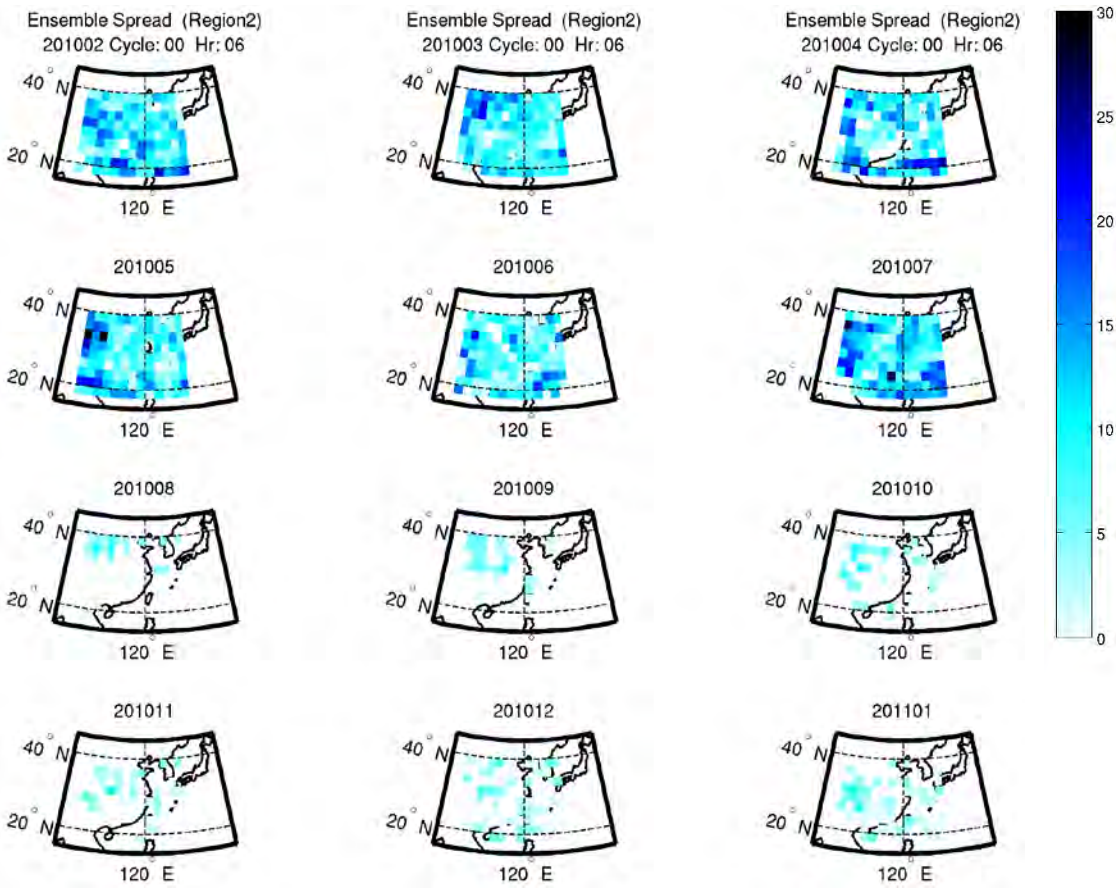

Forecast Percent Correct (Region3)
201002 Cycle: 00 Hr: 06 DT: 20%
Forecast Percent Correct (Region3)
201003 Cycle: 00 Hr: 06 DT: 20%
Forecast Percent Correct (Region3)
201004 Cycle: 00 Hr: 06 DT: 20%
201005
201006
201007
201008
201009
201010
201011
201012
201101
EMean DVote URank WRank
Frequency
p(clear)
0 10 20 30 40 50 60 70 80 90 100

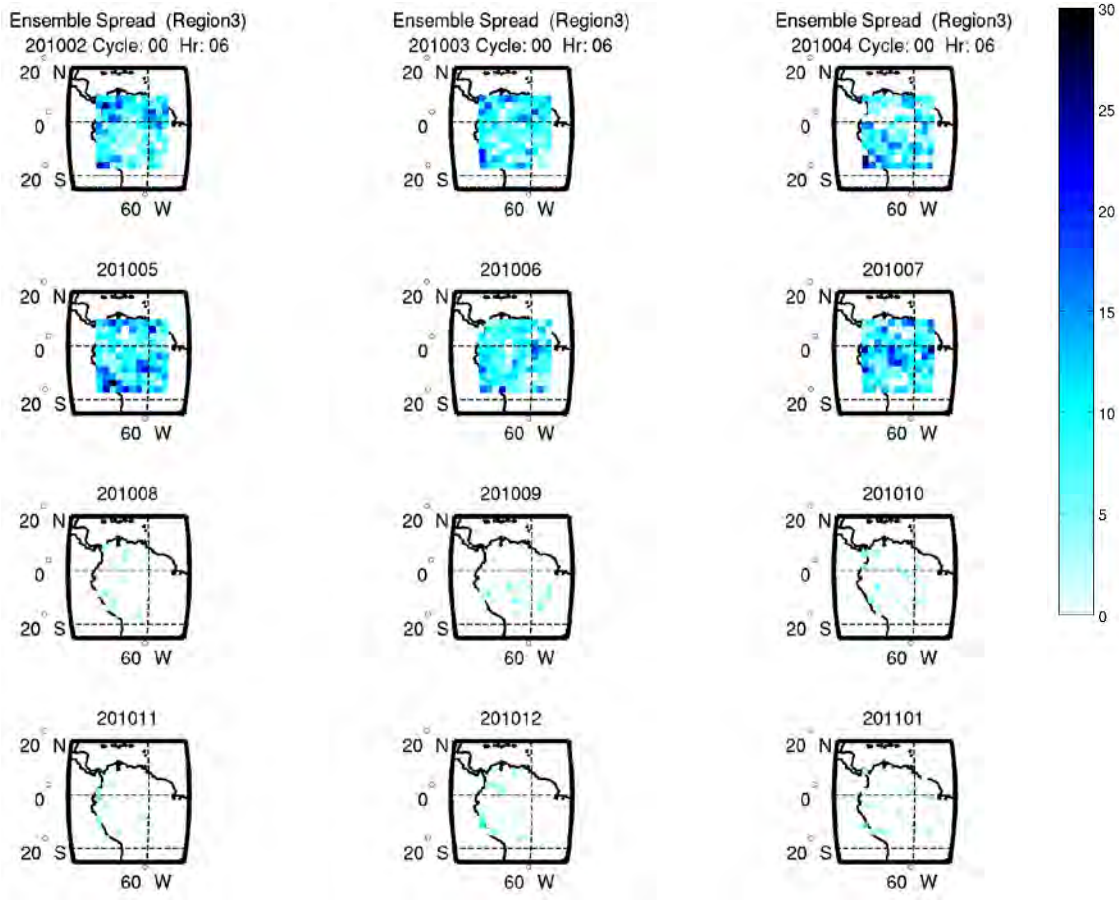

Forecast Percent Correct (Region1)
201002 Cycle: 00 Hr: 06 DT: 20%
Forecast Percent Correct (Region1)
201003 Cycle: 00 Hr: 06 DT: 20%
Forecast Percent Correct (Region1)
201004 Cycle: 00 Hr: 06 DT: 20%
201005
201006
201007
201008
201009
201010
201011
201012
201101
EMean DVote URank WRank
Frequency
p(clear)

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E. ENSEMBLE SPREAD

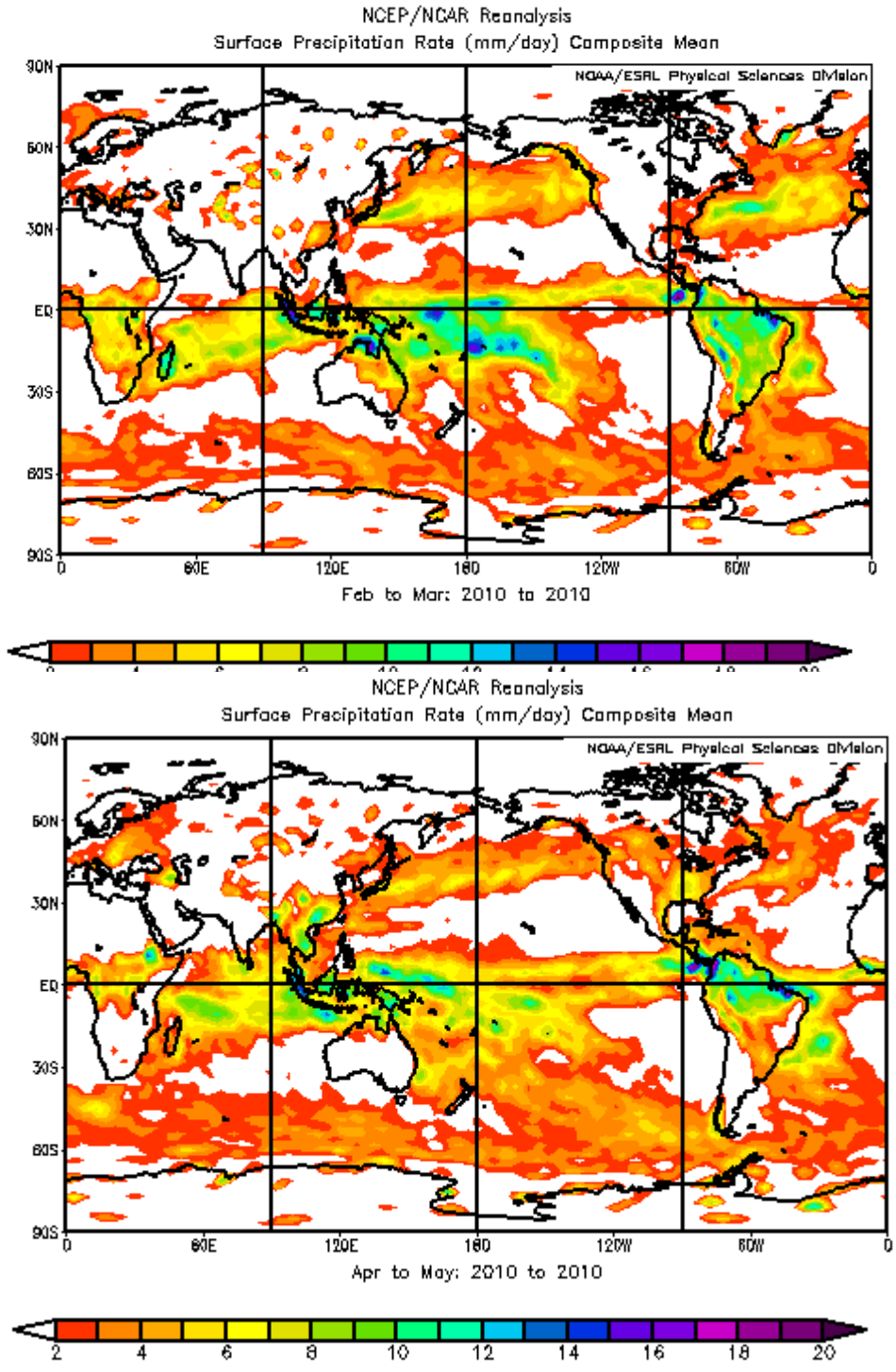


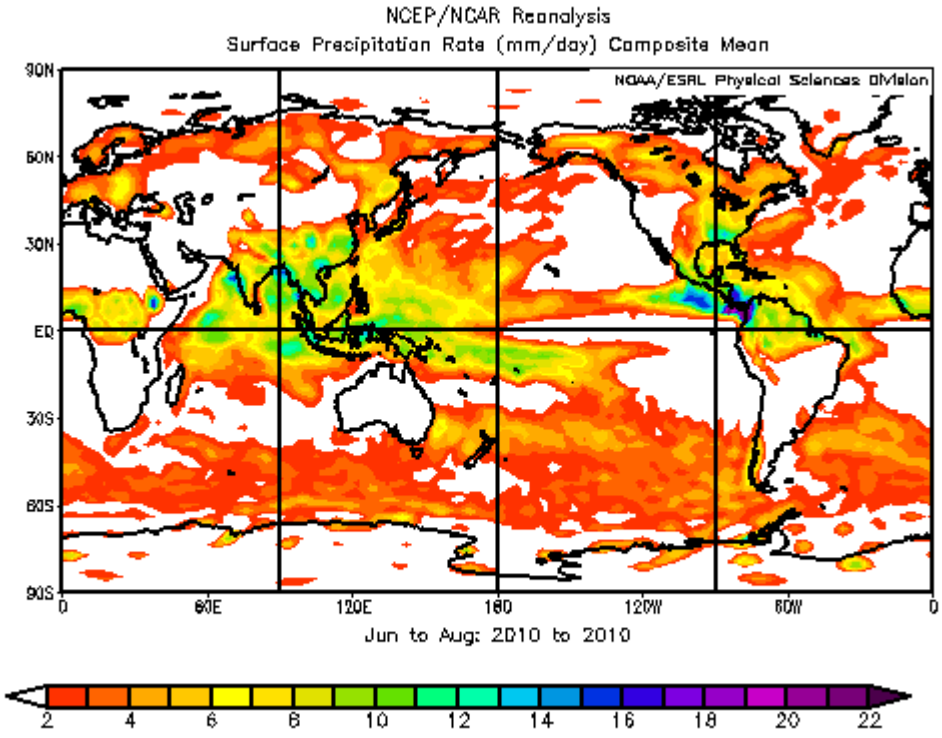

Ensemble Spread (Region2)
201002 Cycle: 00 Hr: 06
Ensemble Spread (Region2)
201003 Cycle: 00 Hr: 06
Ensemble Spread (Region2)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
40° N
20° N
120° E
30
25
20
15
10
5
0

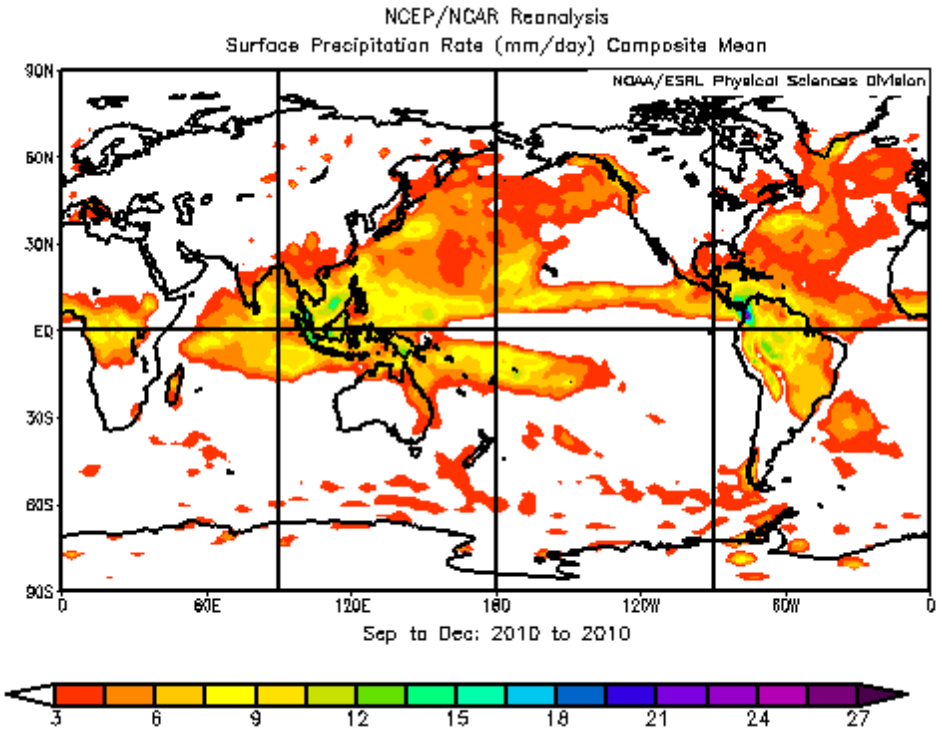

Ensemble Spread (Region3)
201002 Cycle: 00 Hr: 06
Ensemble Spread (Region3)
201003 Cycle: 00 Hr: 06
Ensemble Spread (Region3)
201004 Cycle: 00 Hr: 06
201005
201006
201007
201008
201009
201010
201011
201012
201101
20 N
0
20 S
60 W
30
25
20
15
10
5
0

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX F. NCEP/NCAR MEAN PRECIP REANALYSIS




NCEP/NCAR Reanalysis
Surface Precipitation Rate (mm/day) Composite Mean
NOAA/ESRL Physical Sciences Division
90N
60N
30N
EQ
30S
60S
90S
0
60E
120E
180
120W
60W
0
Jun to Aug: 2010 to 2010
2
4
6
8
10
12
14
16
18
20
22


NCEP/NCAR Reanalysis
Surface Precipitation Rate (mm/day) Composite Mean
NOAA/ESRL Physical Sciences Division
90N
60N
30N
EQ
30S
60S
90S
0
60E
120E
180
120W
60W
0
Sep to Dec: 2010 to 2010
3
6
9
12
15
18
21
24
27

INITIAL DISTRIBUTION LIST

1. Air Force Weather Agency
   Bellevue, Nebraska

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Air Force Weather Technical Library
   Asheville, North Carolina